Estimation de la structure de morceaux de musique par analyse multi-critères et contrainte de régularité

Gabriel Sargent

ANNÉE 2013

UNIVERSITÉ DE RENNES 1

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Traitement du signal

École doctorale Matisse

présentée par

Gabriel SARGENT

préparée à l'unité de recherche IRISA – UMR6074
Institut de Recherche en Informatique et Système Aléatoires
Composante Universitaire

Estimation de la structure des morceaux de musique par analyse multicritère et contrainte de régularité

Thèse soutenue à Rennes le 21 Février 2013

devant le jury composé de :

Gaël RICHARD
Professeur à TELECOM ParisTech, Paris / *Président, Rapporteur*

Myriam DESAINTE-CATHERINE
Professeur à l'ENSEIRB, Bordeaux / *Rapporteuse*

Fabien GOUYON
Chercheur à l'INESC Porto / *Examinateur*

François PACHET
Chercheur à Sony CSL, Paris / *Examinateur*

Frédéric BIMBOT
Directeur de recherche au CNRS / *Directeur de thèse*

Emmanuel VINCENT
Directeur de recherche INRIA / *Co-directeur de thèse*

The Universe is making music all the time.
Tom Waits

Remerciements

Je souhaite tout d'abord remercier Frédéric et Emmanuel pour la qualité de l'encadrement qu'ils m'ont apporté durant ces trois ans sur ce sujet passionnant, ainsi que leur disponibilité et leur bienveillance à mon égard.

Je remercie sincèrement Fabien Gouyon et François Pachet d'avoir accepté la charge d'examinateur, ainsi que Myriam Desainte-Catherine et Gaël Richard pour leurs rapports de thèse d'une grande qualité.

Merci à tous mes collègues de l'équipe METISS, anciens et nouveaux, pour avoir contribué à l'environnement amical et stimulant dans lequel j'ai pu évoluer depuis mon arrivée à Rennes, à Stéphanie pour son dynamisme et son efficacité sur les questions administratives, et à Jean-Christophe, Corentin, Michele, Julien et Dimitris pour m'avoir donné leur avis sur l'écriture et la présentation de ce travail.

Cette thèse fut l'occasion d'un ensemble d'échanges avec la communauté du MIR. Je remercie en particulier Matthew Davies pour m'avoir permis d'utiliser ses programmes d'estimation des temps musicaux et des mesures musicales, Nobutaka Ito, Nobutaka Ono, Stanisław Raczyński, Shigeki Sagayama et les membres du Lab#1 pour leur accueil et l'aide apportée pour la collecte de données symboliques lors de ma visite à l'Université de Tokyo et dans le cadre du projet VERSAMUS. Merci à Benjamin Martin, Matthias Robine et Pierre Hanna pour nos échanges sur la problématique de l'estimation de structure menés à Bordeaux. Merci à Andreas Ehmann, Mert Bay, Kris West et J. Stephen Downie pour leur disponibilité pendant les périodes de rush dans le cadre de MIREX, ainsi qu'à Hélène Lachambre, Maxime Le-Coz, Régine André-Obrecht et Geoffroy Peeters pour l'organisation des évaluations ayant eu lieu dans le cadre de Quaero.

"Last but not least", merci à ma famille et à mes amis qui m'ont aidé à garder les pieds sur terre quand j'en avais besoin : Baptiste, Kevin, J-C, Julie, Emmanuel D., Nadia, Thierry, Marine, les RZ, et, bien sûr, un grand merci à toi, Anne!

Table des matières

Table des figures

Présentation

Les technologies de l'information et de la communication ont aujourd'hui un fort impact sur notre société. Une grande partie de l'humanité est aujourd'hui quotidiennement à leur contact. Ces technologies permettent non seulement l'accès à un large éventail de contenus multimédia numérisés, mais elles vont aussi plus loin en offrant la possibilité aux utilisateurs de créer et de diffuser facilement leurs propres contenus. Il en résulte le développement de bases de plus en plus conséquentes : par exemple le site de vidéo en ligne *Youtube* déclare sur son site que "60 heures de vidéo sont mises en ligne toutes les minutes" [2]. Cette abondance implique le risque que seule une petite partie des données proposées soit finalement utilisée par rapport à sa totalité. Il est ainsi nécessaire de développer des moyens de navigation efficaces à l'intérieur des bases de contenus multimédia à partir de descriptions pertinentes de leurs éléments.

Cette nécessité s'impose également aux bases de morceaux de musique numérisés, qu'il s'agisse de collections personnelles ou de catalogues accessibles par un certain nombre de services en ligne comme *Grooveshark* ou *Spotify*. La taille de ces bases rend la description manuelle très coûteuse, d'où l'intérêt de considérer des moyens de réaliser cette tâche automatiquement. Cette question est à l'origine du développement de la recherche en extraction automatique de contenus musicaux, ou *Music Information Retrieval* (MIR) en anglais. La communauté du MIR réunit notamment des spécialistes de la psychologie, de la musicologie, du traitement du signal et de l'apprentissage automatique.

La musique est un objet complexe et peut être décrite de nombreuses manières, selon son contenu (notes de musique jouées, instruments, nuances...) ou leurs métadonnées (compositeur, date de production, style musical...). Dans cette thèse, nous portons notre attention sur la structure des morceaux de musique en nous focalisant plus particulièrement sur les morceaux diffusés par les média de masse. L'objectif est de décrire l'organisation macroscopique de ces morceaux sous forme d'une suite de segments dont la durée est typiquement de l'ordre de 15 s et qui sont associés à des étiquettes renseignant sur les similarités de leur contenu. Une telle description est utile non seulement à des fins de classement et de navigation, mais aussi pour la création et la recomposition de contenus musicaux (*remixes*) ou pour l'étude et l'analyse musicologique à grande échelle.

La notion de structure n'est cependant pas univoque et peut varier selon les genres et styles musicaux considérés : il est par exemple difficile de décrire un morceau de musique électronique en terme de "couplets" et de "refrains", notions qui s'appliquent habituellement aux morceaux de musique populaire et notamment aux chansons. Cette thèse comporte donc un volet conceptuel visant à définir un type de structure désigné dans cette thèse par le terme de "structure sémiotique" applicable à un large éventail

2. `http://www.youtube.com/t/press_statistics`

de genres différents et dont la spécification et l'annotation ne reposent pas sur des propriétés particulières du contenu musical, ni sur des critères musicologiques postulés *a priori*.

Parallèlement à ces considérations méthodologiques, nous nous penchons sur des méthodes de détection de structure et nous mettons en avant plusieurs approches novatrices pour détecter la structure sémiotique des morceaux. Nous étudions d'une part l'intérêt de favoriser une valeur particulière de la taille des segments recherchés et d'autre part l'utilité de combiner plusieurs propriétés des segments structurels dans le but de localiser automatiquement leurs frontières temporelles. Ces travaux s'appuient sur la conception d'algorithmes et sur un ensemble d'expérimentations inscrites dans le cadre de participations répétées à des campagnes d'évaluation nationales et internationales. Nous explorons également une méthode d'étiquetage des segments utilisant leur organisation interne ainsi qu'une représentation du morceau de musique par automate probabiliste.

Cette thèse s'articule de la manière suivante. Le premier chapitre permet de formuler le problème de l'estimation de la structure des morceaux, son périmètre d'étude et son contexte. Nous présentons ensuite dans le chapitre 2 un panorama des principaux travaux de l'état de l'art relatifs à la spécification et l'estimation de structure. Le chapitre 3 présente les contributions produites pendant la thèse sur la définition et l'annotation de la structure sémiotique, ainsi que les ressources qui en découlent. Le chapitre 4 introduit les approches que nous avons explorées pour l'estimation automatique de la structure musicale. Leur pertinence est évaluée dans le chapitre 5 par l'analyse des résultats des campagnes d'évaluation auxquelles nous avons participé, ainsi que par leur diagnostic complémentaire qui fait l'objet du chapitre 6.

Chapitre 1

La structure musicale : contexte et définitions

Le présent chapitre introduit le contexte et le périmètre de notre étude sur l'estimation automatique de la structure des morceaux de musique. Nous nous intéressons au fait que, bien que souvent mentionnée, la notion de structure d'un morceau connaît un ensemble d'acceptions qui cohabitent aujourd'hui. Ceci découle du caractère complexe de la musique elle-même, qui peut être décrite à la fois selon plusieurs dimensions et selon plusieurs échelles d'étude. Nous présenterons un ensemble d'utilisations intéressantes de la structure dans le cadre de la gestion, de la création et de la compréhension de bases de contenus musicaux.

1.1 Périmètre de l'étude

Du point de vue des sciences de l'ingénieur, la musique peut être considérée comme l'émission d'un ensemble de sons au cours du temps par l'homme et en vue d'être perçus par l'homme. Un morceau de musique peut prendre toute une variété de formes. La multiplicité des genres musicaux qu'il est possible de rencontrer aujourd'hui [1] permet d'en témoigner.

Dans le cadre de cette thèse, on s'intéresse à la musique que l'on se propose de qualifier de "conventionnelle" du point de vue de la culture occidentale. Il s'agit de la musique que l'on rencontre souvent dans les espaces publics, à la radio ou via les autres médias de masse : musique pop, rock, métal, électro, jazz... Les morceaux de musique "conventionnels" présentent une certaine redondance dans leur contenu, afin de susciter des émotions chez ses auditeurs et de faciliter la mémorisation de certains de ses éléments. Ainsi, la musique contemporaine ne fait pas partie de ce cadre.

Nous considérons les morceaux de musique par le biais de leur enregistrement sonore, c'est donc sur la perception de l'audio et non de partitions que se fondera leur analyse. En effet, la musique à laquelle on s'intéresse ici n'est pas forcément issue d'une partition, contrairement à une grande partie de la musique classique. Par ailleurs, il est difficile d'y faire figurer l'ensemble des caractéristiques sonores nécessaires à son exécution. Les annotations relatives à l'expression ou certaines variations de tempo, lorsqu'elles sont présentes, se limitent à quelques expressions de référence laissant place à l'imagination de l'interprète (*ardito, pesante*, ou *accelerando, ritardando* ...).

1. `http://fr.wikipedia.org/wiki/Liste_des_genres_musicaux`

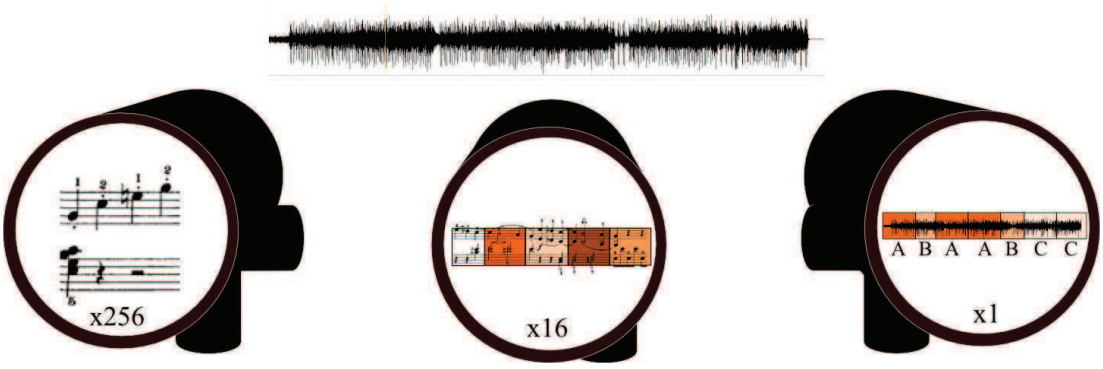

FIGURE 1.1 – Différentes échelles d'étude de la structure

1.2 Attributs de la structure musicale, segmentation et étiquetage

Il est communément admis que l'organisation d'un morceau est pensée (composée) et perçue à plusieurs échelles ([Sny00], p12) (figure 1.1). À court terme, on perçoit des événements sonores fins, tels que les notes de musique, les silences, des sons percussifs, *etc.* Ces événements forment à moyen terme des groupes, que ce soit du point de vue compositionnel, par exemples des motifs ou des cellules musicales [Sir09] ou du point de vue perceptuel, à partir des propriétés qu'ils peuvent partager par proximité temporelle ou fréquentielle, consonance... À long terme, on perçoit des groupes plus longs constitués de ceux du moyen terme. Ces objets sont souvent évoqués par les termes "phrase musicale" [LC00, Sir09], "période" [AdM01], "section musicale" [Got03] ou encore "partie musicale" [PK06].

La diversité de ces dénominations, variables suivant les auteurs et les styles, montre que la notion de structure à long terme d'un morceau de musique peut faire référence à plusieurs objets. Elle reste cependant constituée d'une séquence d'entités que l'on désigne de manière générique par *segments structurels*. Ceux-ci durent en général plus de 10 secondes, et leur contenu est caractérisé par une étiquette. On peut observer l'existence de plusieurs types d'étiquettes : les étiquettes fonctionnelles (introduction, couplet, refrain, solo, *coda* pour les chansons [PK08b], thème, improvisation pour le jazz [Sir09],...), les étiquettes acoustiques, relatives à certaines propriétés du contenu musical (*lead* : chant, solo de guitare, volume : *fade-in*, *fade-out*) [PD09], ou des étiquettes relatives à une tradition d'écriture particulière (par exemple en musique classique : les formes rondo et sonate [AdM10]).

1.3 Positionnement

La tâche d'estimation automatique de la structure des morceaux de musique s'inscrit dans le domaine de la recherche d'informations dans les contenus musicaux, ou *Music Information Retrieval* (MIR). Il s'agit d'un domaine relativement récent qui a commencé à se développer il y a une dizaine d'années environ [DBC09]. Au sein de celui-ci, les disciplines liées aux sciences de l'ingénieur, à la musicologie et aux sciences cognitives interagissent pour la recherche de méthodes et d'algorithmes d'extraction de contenus musicaux à des fins de classement, de compréhension ou de réutilisation de ces contenus

[CVG+08].

Ce domaine comprend par exemple les tâches de description automatique de musique en vue de leur classement (par genre, par émotions...) et de leur recherche par requêtes spécifiques (fredonnement, mots clés...). Il comprend aussi la transcription automatique des notes ou des accords d'un enregistrement audio, ou la séparation et le remixage des instruments de musique qui y sont présents. Une présentation plus détaillée de ce domaine et de ses enjeux est disponible dans [Ori06].

L'estimation de la structure musicale suscite un intérêt croissant depuis quelques années comme le montrent les *reviews* menées dans [PMK10, KD06, DG08]. Un ensemble de campagnes d'évaluation et de projets incluent cette thématique. La campagne MIREX [2], d'envergure mondiale, comprend une tâche de segmentation structurelle (*structural segmentation*) depuis 2009 [EBD+11]. Le projet européen Quaero [3] sur le traitement automatique de contenus multimédia comporte un thème de recherche lié à l'estimation de la structure et au résumé musical (*music structuring and summarization*) et mène une campagne d'évaluation annuelle sur le sujet, également depuis 2009 [Pee11, LA11, SBV10a]. Plus récemment, le projet anglo-américain SALAMI [4] [SBF+11] se concentre exclusivement sur l'analyse structurelle automatique d'une grande base de morceaux de musique.

1.4 Applications

Depuis plusieurs années, les techniques d'encodage de flux audio et l'amélioration des capacités de stockage permettent la création et le développement de grandes bases de morceaux de musique, personnelles ou partagées via des services particuliers, par exemple Grooveshark [5], ou Spotify [6]. L'analyse automatique de la structure des morceaux peut-être utile pour un certain nombre de tâches relatives à ces bases, que l'on se propose de regrouper en trois familles comme illustré dans la figure 1.2 :

Gestion de bases de contenus musicaux La structure peut être utilisée dans un but d'indexation des morceaux de musique pour permettre un accès efficace et rapide au contenu des bases [GSMA12], et garder une certaine visibilité de leur ensemble. La structure est une caractéristique macroscopique d'un morceau de musique. Elle peut être utile afin d'améliorer les méthodes de recherche de morceaux similaires, et en particulier des morceaux repris par plusieurs artistes (ou *cover songs* [EP07]). Connaître la structure d'un morceau permet de naviguer efficacement à l'intérieur d'un morceau. La connaissance des segments structurels et de leur classement permet l'accès rapide à la portion de morceau qui intéresse l'utilisateur d'une base, ou l'ingénieur du son lors d'un travail de production sonore. Enfin, elle donne la possibilité de produire des aperçus sonores permettant de se faire une idée rapide du contenu des morceaux. Un aperçu peut être une portion musicale "qui représente le mieux le morceau" (ou *thumbnail* [MGJ11]), ou un résumé musical résultant de la combinaison d'une version raccourcie des segments importants du morceau (typiquement introduction, couplets et refrains dans le cas d'une chanson) [PLR02].

2. www.music-ir.org/mirex/
3. www.quaero.org/
4. http ://www.music-ir.org/ ?q=node/14
5. `http://grooveshark.com/`
6. `http://www.spotify.com/`

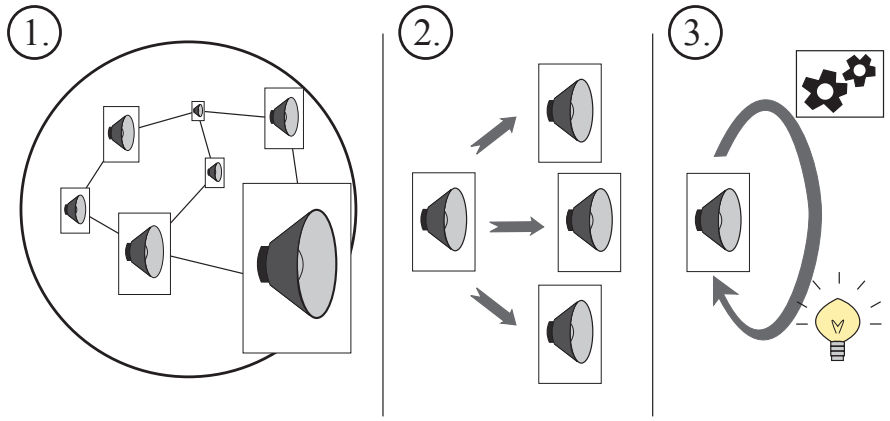

FIGURE 1.2 – Trois familles d'applications liées à l'estimation de structure des morceaux de musique : la gestion de bases de contenus musicaux (1), la création et l'analyse automatique de contenus musicaux (2), et la compréhension de contenus musicaux (3).

Création et analyse automatique de contenus musicaux La structure peut être utile à la génération automatique de "remix" ou de bande-son pour une production multimédia. Il peut s'agir de recomposer un morceau de manière à ce que sa durée totale corresponde aux besoins de l'utilisateur (séquence de film, reportage...), ou encore de créer la bande-son d'un jeu vidéo à partir de l'un des morceaux favoris du joueur.

L'introduction d'informations structurelles peut constituer une aide aux outils d'analyse et de traitement de la musique, que ce soit pour la séparation de sources sonores [LRB+12], l'estimation d'accords [MND09], ou l'estimation du tempo [Dan05].

Compréhension de contenus musicaux Dans le domaine de la musicologie, la structure pourrait constituer un outil pour la caractérisation des genres musicaux par l'étude statistique de grandes bases de morceaux [TC02]. Il serait aussi intéressant de considérer la structure afin d'analyser les mécanismes de perception et d'apprentissage de la musique [BMTP99]. On peut lier la stratégie d'exposition des segments structurels à la volonté du compositeur de susciter des émotions chez d'auditeur ainsi que de mémoriser certains passages [Slo91]. Enfin, elle peut constituer une aide à l'apprentissage et à l'interprétation musicale. La possibilité de repérer l'organisation à long terme d'un morceau rend plus aisée sa compréhension et son assimilation.

Ainsi, la structure peut contribuer à un large champ d'applications liées à la gestion, l'expansion et la compréhension des bases de morceaux de musique.

1.5 La structure musicale : un objet complexe

Les étiquettes évoquées dans la partie 1.2 renvoient à des notions qui ne sont pas clairement définies ou trop spécifiques à un genre musical. Ainsi, nous écarterons par la suite ces notions *sémantiques* pour nous concentrer sur une étude *sémiotique* de la structure, c'est-à-dire fondé sur l'analyse du contenu musical de ses différentes portions [7].

7. Nous désignons dans notre travail une portion comme un segment quelconque du morceau.

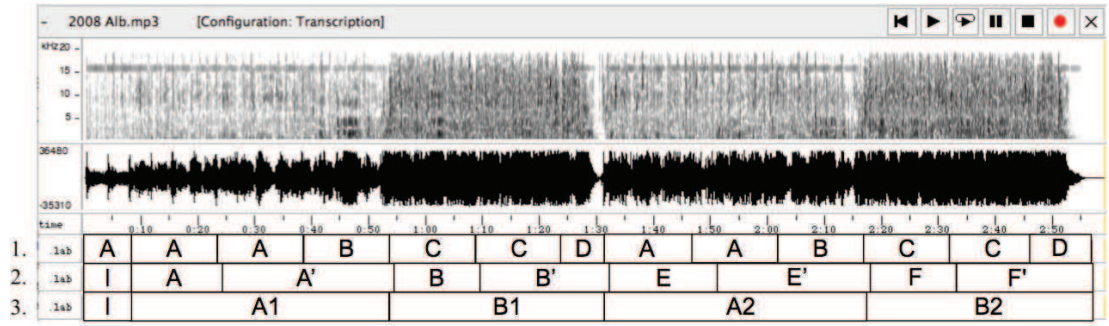

FIGURE 1.3 – Trois exemples d'annotations structurelles du morceau *Zemrën lamë peng*, de Olta Boka. Des différences peuvent être observées au niveau des frontières structurelles et des étiquettes bien qu'elles ne contredisent pas la caractérisation de la structure proposée dans la partie 1.2. Cette visualisation est obtenue par le logiciel Wavesurfer [9].

La structure que l'on considère est caractérisée par une séquence de segments structurels. Chaque segment structurel est défini par :

- un *instant de début*,
- un *instant de fin*,
- une *étiquette* attribuée selon la *similarité* de son contenu avec celui des autres segments du morceau.

Chaque étiquette, issue d'un alphabet arbitrairement choisi (lettres, numéros,...), réfère à un groupe de segments structurels similaires, ou *classe d'équivalence*.

La tâche d'estimation de la structure d'un morceau consiste à analyser le contenu musical afin de réduire le morceau à une séquence de segments structurels. Elle est donc constituée de deux étapes, qui peuvent être exécutées successivement ou conjointement :

- la *segmentation* : l'estimation des frontières des segments structurels,
- l'*étiquetage* : le regroupement des segments structurels en classes.

Cependant, cette caractérisation ne conduit pas à définir un objet unique pour chaque morceau : proposez à plusieurs personnes d'annoter une même chanson, et il est fort probable que les annotations produites ne coïncident pas en tout point. Peiszer a pu souligner les différences entre les annotations de référence de plusieurs morceaux de différents laboratoires de recherche pour l'estimation de structure [Pei07].

1.5.1 Incohérence et ambiguïté

Considérons par exemple la chanson représentant l'Albanie dans le concours Eurovision 2008 (*Zemrën lamë peng* de Olta Boka), dont la forme d'onde et le spectrogramme sont représentés dans la figure 1.3. Ce morceau de musique pop semble de prime abord avoir une structure assez simple : on remarque rapidement, après une brève introduction, une alternance de couplets et de refrains correspondant à deux régimes stationnaires différents (troisième annotation du panneau du bas). La densité sonore des refrains est notamment plus forte que celle des couplets, comme nous pouvons le visualiser sur son spectrogramme (panneau du haut) et sa forme d'onde (panneau du milieu). Trois annotations structurelles sont représentées dans le panneau du bas. On y observe plusieurs différences concernant les frontières et les étiquettes structurelles,

9. http ://www.speech.kth.se/wavesurfer/

ainsi que l'échelle d'étude. Cependant, ces annotations sont toutes acceptables selon la caractérisation de la structure décrite dans la partie 1.2.

Du point de vue des frontières structurelles, l'annotation numéro 1 est synchronisée aux cycles de la partie instrumentale. Ici, un cycle correspond à une répétition d'un motif de guitare sur laquelle se synchronisent la guitare basse et les percussions. La numéro 2 utilise la mélodie chantée, qui démarre en moyenne une seconde plus tard. On observe ainsi un décalage d'environ une seconde pour la plupart des frontières. La mélodie chantée dans les segments structurels de l'annotation 2 ayant les étiquettes A', C', E' et F' évolue sur une durée plus longue que les segments A, C, E et F. Ceux-ci tendent à correspondre aux segments issus de la fusion de A et B, puis C et D qui se succèdent dans l'annotation 1, qui privilégie les classes de segments structurels ayant des durées comparables. Notons enfin que l'annotation 3, basée sur l'homogénéité de la densité sonore est composée de segments nettement plus longs, ce qui correspond à une échelle d'étude moins fine que les deux autres annotations.

Du point de vue des étiquettes structurelles, l'annotation 1 propose d'étiqueter A le premier segment car son contenu musical correspond à la partie instrumentale des autres segments notés A. L'annotation 2 le considère à part avec l'étiquette I car il ne contient pas de mélodie chantée. Enfin, l'annotation 1 considère que deux segments semblables à un changement de tonalité près appartiennent à une même classe (segments A, B, C), contrairement à l'annotation 2 (qui considère de nouvelles classes E, E', F et F').

On remarque ainsi que l'annotation de la structure va dépendre d'un ensemble de choix relatifs :

- aux caractéristiques musicales jugées pertinentes pour l'étude de la structure,
- à l'échelle d'étude de la structure (fine, grossière),
- à la notion de similarité entre segments.

Ces choix, qui peuvent varier au cours d'un même morceau, influencent à la fois la position des frontières et les étiquettes des segments structurels. L'absence de spécification précise et commune de la structure recherchée implique une certaine incohérence des annotations produites. La question de la spécification de la structure musicale, nécessaire afin d'assurer la cohérence des annotations, est étudiée dans le cadre du chapitre 3 de cette thèse.

Une fois la structure spécifiée, on peut noter que certains aspects de la structure de certains morceaux peuvent rester ambigus. Considérons par exemple le morceau *Brain Damage* du groupe Pink Floyd dont le spectrogramme, la forme d'onde puis l'annotation de référence produite selon la méthodologie introduite dans la partie 3 sont représentés dans la figure 1.4. Les segments étiquetés 2 correspondent à un segment de transition (un *break* de batterie) situé entre un segment A (un couplet) et un segment B (un refrain). La taille de ces segments est petite devant celle des autres. Il n'a pas été rattaché à l'un de ses voisins selon les annotateurs qui n'ont pu décider s'il partageait plus de relations avec l'un ou l'autre. Pourtant, les segments $[A + 2]$ et $[2 + B]$ respectivement issus de la fusion de A avec 2 puis 2 avec B semblent être des segments structurels acceptables. Nous ne nous attacherons pas à résoudre les problèmes de cet ordre dans cette thèse.

1.5.2 La musique : un objet multidimensionnel et multi-échelles

L'existence de plusieurs annotations d'un même morceau est liée au fait que la musique est un objet complexe.

Il s'agit d'un objet *multidimensionnel*, dans le sens où elle peut être décrite selon

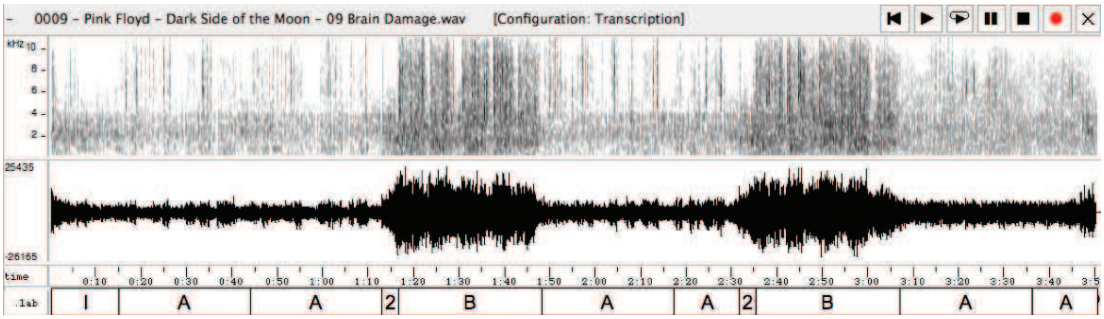

FIGURE 1.4 – Spectrogramme (haut), forme d'onde (milieu) et annotation de la structure musicale de référence (bas) du morceau *Brain Damage* de Pink Floyd.

plusieurs caractéristiques que nous désignerons dans la suite de cette thèse sous le terme de *strates d'information musicales*. La musique que l'on considère dans cette étude est fortement influencée par le système tonal. On peut référencer huit strates musicales principales pour la description du contenu d'un morceau de musique, dont celles habituellement considérées dans le cadre de l'estimation de structure [PMK10] :

- l'harmonie (succession des accords au cours du temps, où un accord résulte de la superposition d'au moins trois sons [AdM01]),
- l'intensité sonore (amplitude du son, niveau sonore résultant de l'exécution du morceau),
- la mélodie (succession ordonnée de sons musicaux, articulée à partir de rythmes et de hauteurs [AdM01]),
- les paroles (organisation des mots, rimes, ...),
- le rythme (résultat de l'organisation des durées, des timbres ou des accents successifs des sons musicaux émis au cours du morceau),
- le tempo (vitesse d'exécution du morceau),
- le timbre (résultant des timbres des instruments présents [Sir09]),
- la tonalité/modalité (ensemble de hauteurs obtenues à l'aide d'une note de référence, la tonique, et d'un ensemble d'intervalles musicaux, le mode [Sir09]).

La musique est aussi un objet *multi-échelles* car elle affiche plusieurs échelles d'organisation temporelle, comme nous l'avons décrit en 1.2. De ces deux caractéristiques découlent un réseau de relations complexe qui rend ambiguë la notion de "structure à long terme" couramment utilisée.

Cette complexité implique que la structure telle que caractérisée dans la plupart des travaux s'intéressant à son extraction n'est pas unique. Ceci pose donc évidemment problème lorsque l'on souhaite l'estimer automatiquement.

1.6 Résumé du chapitre et objectifs de la thèse

Le présent chapitre nous a permis de définir le périmètre de notre étude sur l'estimation de la structure des morceaux de musique, et de le contextualiser par rapport au domaine des sciences de l'ingénieur au sein du domaine MIR. Ce sujet est tout d'abord intéressant d'un point de vue scientifique car il se focalise sur la description de l'organisation d'un objet complexe : la musique est un flux d'information multi-strates, et son organisation peut être observée à plusieurs échelles. À cela, ajoutons le fait que la notion de ressemblance entre deux portions d'un morceau dépend de leur contenu

musical et peut être influencé par leur contexte dans ce morceau. De fait, le problème de l'estimation de structure est un problème mal posé, car la notion même de structure musicale (à long terme) n'est pas clairement définie et peut être annotée différemment suivant différents annotateurs. Ce sujet est ensuite prometteur d'un point de vue applicatif. La musique numérisée fait aujourd'hui partie du quotidien des hommes et un ensemble de services proposent l'accès à de grandes bases de contenus musicaux. La structure peut contribuer à la gestion, la création, l'analyse et la compréhension de telles bases.

Dans cette thèse nous nous attachons au problème de l'estimation de la *structure sémiotique* des morceaux de musique "conventionnels" à partir du signal audio. Les aspects multi-strates et multi-échelles de la musique font que nous considérons en particulier :

- l'utilisation conjointe de plusieurs caractérisations du contenu musical des segments structurels (approche multicritère),
- l'introduction d'une contrainte sur la taille de ces segments (contrainte de régularité),
- l'utilisation d'une nouvelle caractérisation de l'organisation interne des segments structurels : le modèle système-contraste,

dans le cadre de l'estimation des frontières structurelles. Le problème de l'étiquetage sémiotique est ensuite considéré par l'étude d'un système expérimental d'étiquetage sémiotique basé sur la sélection d'automates probabilistes et utilisant le modèle système-contraste.

La thèse suit le plan suivant. Le chapitre 2 est l'occasion d'une part de nous pencher sur les travaux ayant pour objet la spécification et l'analyse manuelle de la structure d'un morceau de musique, puis, d'autre part, d'étudier les approches et techniques de l'état de l'art utilisées pour l'estimation de structure, partielle ou complète. Ceci nous permettra de mettre en lumière l'absence de consensus quand à la notion de structure musicale, ainsi que plusieurs axes de recherche prometteurs pour son estimation.

Le chapitre 3 porte sur la spécification de la structure sémiotique que nous cherchons à estimer dans le cadre de cette thèse. Nous posons le problème dans le cadre du structuralisme en linguistique (domaine qui s'est concentré sur l'étude des mécanismes du langage humain) et nous proposons une spécification et une méthodologie d'annotation basées sur l'expérience d'écoute des annotateurs et avec l'objectif de couvrir un vaste ensemble de genres et styles musicaux.

Suite à l'étude des techniques de l'état de l'art et sous l'influence de ces avancements méthodologiques, nous introduisons dans le chapitre 4 un cadre formel et plusieurs approches novatrices pour l'estimation des frontières structurelles. Nous nous concentrons ensuite sur la problématique de l'étiquetage sémiotique par le biais de plusieurs configurations d'un système de sélection d'automates, en supposant les frontières structurelles connues.

Le chapitre 5 présente et analyse les performances de plusieurs systèmes d'estimation de structure basés sur les approches du chapitre 4 et évalués dans le cadre de campagnes évaluations d'envergure internationale et nationale en 2010, 2011 et 2012. Ceci permet de situer l'efficacité de nos approches par rapport à l'état de l'art, et d'observer le comportement des différents systèmes sur les différentes bases d'annotations pour l'évaluation. Une partie de ces annotations est issue de la méthodologie du chapitre 3.

Enfin, le chapitre 6 porte sur le diagnostic des modules d'estimation des frontières

structurelles et des étiquettes sémiotiques fondés sur nos approches. Nous utilisons la base RWC *Popular Music* composée de 100 morceaux de musique pop japonaise fortement inspirée de musique pop rock occidentale pour l'estimation des frontières, et une base constituée de 100 morceaux de pop occidentale, sélectionnés parmi les meilleures ventes et annotés dans le cadre du projet Quaero. Nous étudions l'intérêt d'une approche multicritère et de la contrainte de régularité pour l'estimation des frontières, ainsi que l'intérêt de considérer le modèle système-contraste pour l'étiquetage.

Le fait de présenter les performances de systèmes mettant en oeuvre nos approches dans le cadre de campagnes d'évaluation avant d'étudier leurs différents modules peut sembler assez atypique. Nous justifions ce choix par le fait que la participation à ces campagnes d'évaluation annuelles s'est avérée être une source de motivation importante influençant à chaque fois notre progression dans cette thèse, par la perspective concrète de comparer nos approches à celles de l'état de l'art à l'aide de systèmes complets et opérationnels. Nous reconnaissons qu'une telle approche a ses limites : outre la nécessité de se pencher sur certaines questions pratiques (comme le respect des formats d'entrée et de sortie imposés), son caractère exploratoire implique que les systèmes soumis peuvent ne pas avoir atteint un stade de maturité optimal. C'est pour cela que nous avons effectué dans le chapitre 6 le diagnostic de plusieurs modules constituant ces systèmes, dont certains ont été améliorés.

Chapitre 2

État de l'art

Comme le souligne Paulus, "la tâche d'*analyse structurelle* renvoie à un ensemble de problèmes, et différents chercheurs ont poursuivi des buts légèrement différents dans ce contexte. Ils partagent cependant approximativement la même échelle temporelle d'analyse" [PMK10].

Le terme d'analyse structurelle est ainsi à prendre au sens large, et concerne autant les méthodes d'estimation d'un aspect structurel particulier comme la segmentation seule [Jen07], l'extraction de vignettes sonores (*thumbnails*) [LC00] ou de refrains [Got03], que les méthodes d'estimation de la "structure complète", c'est-à-dire la segmentation et l'étiquetage de l'ensemble du morceau.

Ce chapitre décrit tout d'abord les principaux travaux relatifs à la spécification de la structure en musicologie et en MIR (partie 2.1), pour ensuite présenter les différentes méthodes d'analyse structurelle selon une grille d'analyse personnelle (partie 2.2). Plutôt que de classer les différents systèmes de l'état de l'art suivant leur but ou suivant les outils utilisés [PMK10], nous mettons en valeur les hypothèses et les choix effectués pour les construire selon trois axes : les strates d'information musicales considérées pour l'analyse structurelle (partie 2.3), la manière de caractériser les propriétés du contenu musical des segments structurels (*critères audio*, partie 2.4), et les hypothèses sur la structure musicale recherchée (*contraintes structurelles*, partie 2.5).

2.1 Travaux relatifs à la spécification de la structure musicale

2.1.1 Musicologie

On peut se demander si la musicologie ne pourrait pas offrir certains outils nous permettant de limiter l'ambiguité de la notion de structure musicale. Une étude exhaustive de la littérature musicologique sur le sujet sort du cadre de notre étude. Bent et Drabkin [BD98, ch. 4] répertorient notamment un ensemble d'approches pour modéliser la musique et plusieurs procédés d'analyse. Ces approches sont pour la plupart spécifiques à la musique classique, et se concentrent sur la tonalité et des cadences particulières [1], ce qui est contraignant pour le type de musique que l'on considère.

Parmi les références musicologiques citées en MIR, Peiszer se fonde sur les travaux de

1. Il s'agit d'une "formule mélodique ou harmonique qui ponctue ou conclut une phrase ou une oeuvre" [AdM01].

Middleton [Mid90] afin de discuter des différents procédés d'analyse structurelle [Pei07, pp. 12–14]. Selon lui, il est difficile de comparer deux descriptions d'un même morceau qui sont issues de deux procédés différents, ce qui rend les choses assez compliquées pour un non-musicologue. Il souligne néanmoins la pertinence de l'*analyse paradigmatique* [Ruw87], qui consiste à segmenter un morceau selon ses répétitions. En conséquence, les segments obtenus à une échelle donnée ont des tailles proches. Cette approche est notamment reprise dans les travaux de Paulus et Klapuri [PK06].

2.1.2 MIR

Dans le cadre du MIR, un grand nombre de travaux faisant intervenir une analyse structurelle des morceaux de musique se focalisent sur un aspect applicatif : génération de résumé [LC00], de vignette sonore [MGJ11], ou encore extraction du refrain [Got03]. Cela permet de simplifier le problème en extrayant un aspect spécifique de la structure, et tend à expliquer le peu de références explicites à des notions musicologiques sur la structure.

Peeters et Deruty proposent une caractérisation multidimensionnelle de la structure dans [PD09], partiellement reprise par la suite dans [SBF+11]. Il s'agit d'annoter la structure suivant plusieurs points de vue indépendants : les *phrases musicales* (progressions d'accords), l'arrangement (richesse de l'instrumentation), les rôles instrumentaux (instrument principal à l'écoute) et la fonction (intro, transition, refrain, solo, coda...). Les auteurs proposent de synchroniser les frontières sur l'échelle des premiers temps des mesures [2] (ou *downbeats* en anglais). L'ambiguïté d'échelle est prise en compte en offrant la possibilité d'annoter la similarité à l'échelle inférieure, lorsqu'un segment peut se décomposer en deux sous-segments.

Bruderer *et al.* visent à caractériser les frontières des segments de manière statistique [BMK06]. Il s'agit de conserver celles qui auront été le plus annotées par un ensemble de personnes. Cette étude souligne que les annotations coïncident souvent avec un changement significatif et durable de la composition instrumentale au cours du morceau (rupture de timbre).

Ce bref panorama des domaines de la musicologie et du MIR montre qu'il n'y a pas, aujourd'hui, de convention unanimement adoptée par la communauté pour l'estimation de la structure des morceaux. Ceci nous a amené à participer à un ensemble de travaux visant à spécifier une structure mono-dimensionnelle issue d'une analyse multi-strates qui soit applicable à un large éventail de genres musicaux et telle que sa méthodologie d'annotation soit la plus simple et reproductible possible. Nous faisons dans le chapitre 3 une synthèse des publications issus de ces travaux.

2.2 Vue d'ensemble d'un système d'estimation de structure, notions de critères et de contraintes

Dans la partie 1.2, nous avons évoqué qu'un système d'estimation de la structure des morceaux de musique est composé de deux grandes étapes (figure 2.1). Dans une première étape, un ensemble de descripteurs est extrait du signal audio, qui permet

2. Une mesure correspond à un groupe de temps musicaux successifs. Il existe, selon ([AdM01],p. 550), une "hiérarchisation des temps successifs de la mesure depuis l'époque baroque ; le premier temps est considéré comme *fort* c'est-à-dire naturellement accentué, les autres peuvent être soit *faibles*, non accentués, soit *demi-forts*, d'une accentuation plus légère que le temps fort."

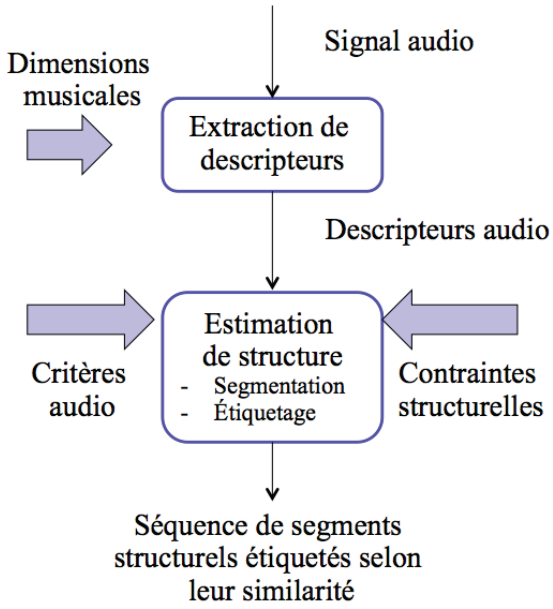

FIGURE 2.1 – Principales étapes et composantes d'un système d'estimation de la structure.

de mettre en valeur certaines propriétés musicales. Dans une deuxième étape, ceux-ci sont utilisés pour l'estimation de la structure du morceau considéré. Cette étape est constituée de deux sous-étapes : le morceau est découpé en segments structurels (segmentation), qui sont regroupés en classes en fonction de leur similarité (étiquetage). Certaines méthodes reviennent sur la segmentation après l'étape d'étiquetage, par exemple en fusionnant les segments voisins de même classe [PK06].

La majorité des systèmes d'analyse structurelle est composée de ces étapes. Cependant, pour chacune d'entre elles, on doit effectuer un ensemble de choix afin de répondre aux questions suivantes :

- choix d'une ou plusieurs strates musicales : à quelles propriétés musicales nous intéressons nous ?
- choix d'un ou plusieurs critères audio : comment caractérise-t-on un segment structurel ?
- choix de contraintes structurelles : quelles hypothèses fait-on sur la structure du morceau afin de limiter le nombre de structures visées ?

La figure 2.1 montre à quelle étape intervient chacun de ces trois choix. Nous décrivons maintenant l'état de l'art selon ces trois axes dans les trois parties suivantes.

2.3 Principaux descripteurs pour l'estimation de structure

Nous avons répertorié un ensemble de strates musicales dans la partie 1.2. Les travaux actuels utilisent principalement le timbre, l'harmonie, la mélodie et le rythme pour l'analyse structurelle automatique. Afin de les étudier au cours du morceau de

musique, ce dernier est divisé en petites trames [3] de quelques dizaines de millisecondes à partir desquelles sont extraits un ou plusieurs descripteurs. Ces trames peuvent se recouvrir temporellement, le plus souvent de moitié.

Descripteurs numériques, descripteurs symboliques On peut classer les descripteurs en deux types :

- Les descripteurs de type *numérique*. Il s'agit de toute valeur numérique extraite du signal audio, par exemple les puissances associées aux différentes fréquences composant le signal à un instant donné.
- Les descripteurs de type *symbolique*. Il s'agit de symboles parmi un alphabet fini qui caractérisent le contenu des portions de signal auxquels ils sont rattachés. Ces descripteurs peuvent faire référence à des concepts de plus haut niveau comme les accords, ou être obtenus par quantification vectorielle de descripteurs numériques.

De rares travaux portent sur la recherche de structure à partir d'une représentation symbolique des morceaux de musique par l'analyse de partitions ou de fichiers MIDI [ANO06]. Ces représentations peuvent être inaccessibles, et un grand nombre de morceaux de musique "conventionnels" d'aujourd'hui n'est pas issu d'une partition écrite. Cependant une partie du domaine du MIR s'intéresse à leur estimation à partir de l'audio. Il est ainsi possible d'accéder à un ensemble de descriptions symboliques (mélodie [PE05], accords [UUN+10], tonalité [PP09]...) qui sont aujourd'hui principalement utilisées pour la classification automatique de morceaux selon leur ressemblance, leur genre musical, l'humeur qu'ils suscitent, *etc* [TWV05]. Il n'existe à notre connaissance pas de travaux visant à estimer la structure sémiotique d'un morceau à partir de ces estimations. Nous utilisons des estimations d'accords dans le cadre d'un système d'estimation de structure dans le chapitre 4, et des vecteurs de chroma quantifiés pour l'étiquetage des segments structurels dans le chapitre 5.

MFCCs en tant que descripteurs de timbre De nombreuses méthodes d'estimation de structure considèrent l'évolution des caractéristiques timbrales des morceaux de musique au cours du temps [PMK10], par l'intermédiaire de descripteurs dits "de timbre". Une façon de décrire le timbre consiste à encoder l'enveloppe spectrale de la portion de signal contenue dans la trame qui lui est associée. La durée d'une trame varie typiquement entre quelques dizaines à quelques centaines de millisecondes (par exemple 25 ms [LC00] ou 400 ms [KS10]).

Les descripteurs les plus utilisés dans ce cadre sont les coefficients cepstraux à l'échelle de Mel, ou MFCCs (*Mel Frequency Cepstral Coefficients* [RJ93]). Les MFCCs d'une portion de signal x sont obtenus de la manière suivante [LC00] :

1. on calcule le spectre en amplitude de x par une Transformée de Fourier Discrète,
2. on calcule le logarithme du spectre en amplitude,
3. ce log-spectre est filtré par un banc de filtres triangulaires régulièrement espacés sur l'échelle de Mel,
4. les coefficients cepstraux sont obtenus en effectuant une Transformée en Cosinus Discrète.

3. Nous utilisons le terme trame pour désigner, de manière générique, des fenêtres temporelles dont la durée peut être fixée ou varier à l'échelle du morceau. Nous verrons par exemple que les trames peuvent dépendre des temps musicaux, dont la période peut varier au cours du temps.

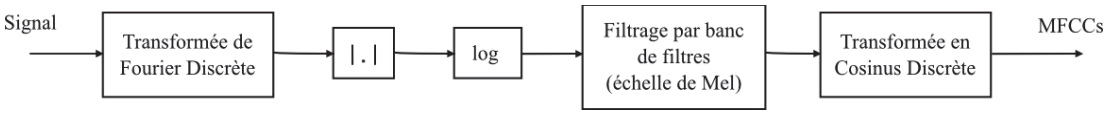

FIGURE 2.2 – Processus de génération des coefficients MFCC.

Si l'on suppose que le signal étudié est produit par un modèle filtre/excitation [Fan60], le fait d'utiliser le logarithme du spectre comme dans l'étape 2. permet de décorréler l'excitation et l'enveloppe. Ainsi, les premiers coefficients MFCC vont encoder l'enveloppe spectrale et les autres la structure fine du spectre.

L'échelle de Mel utilisée dans l'étape 3, dérivée du logarithme, permet de prendre en compte certaines caractéristiques de l'audition humaine. L'échelle linéaire et l'échelle de Mel sont liées par la formule 2.1 [KD06, p. 26], dans laquelle f_{mel} est en Mel et f est en Hz :

$$f_{\text{mel}} = 2595 \log_{10} \left(\frac{f}{700} + 1 \right). \tag{2.1}$$

La Transformée en Cosinus Discrète de l'étape 4 permet d'obtenir un petit nombre de coefficients aussi décorrelés que possible, dans l'optique d'obtenir une description compacte et informative du contenu timbral du signal.

Les MFCCs ont initialement été proposés pour le traitement automatique de la parole [RJ93], en particulier pour la reconnaissance de phonèmes. En musique, on s'intéresse aux premiers coefficients, typiquement jusqu'au treizième ou au vingtième, pour représenter l'enveloppe spectrale globale du signal à un instant donné.

D'autres descripteurs de timbre existent, mais ceux utilisés dans le cadre de l'estimation de la structure reprennent peu ou prou ce processus avec quelques variantes. Ils peuvent utiliser une échelle de fréquences linéaire [TSB05] ou par octaves [Mad06] à la place de l'échelle de Mel, ou une analyse en composantes principales, méthode de réduction de la dimension des données par l'identification des axes de plus forte variance, à la place de la Transformée en Cosinus Discrète (descripteur *AudioSpectrumProjection* issu de la norme MPEG-7 [LS08]).

Certaines approches modélisent les vecteurs MFCC par leurs moments (barycentre spectral, écart-type, asymétrie, *kurtosis*...) mais ceux-ci sont moins bien adaptés que les coefficients eux-mêmes pour l'analyse structurelle selon [AS01].

Vecteurs de chroma et autres descripteurs de type tonal Les descripteurs de type "tonal" visent à décrire le contenu d'une portion de signal en terme d'un ensemble de hauteurs tonales en référence à la gamme chromatique de la théorie de la musique occidentale. Ils sont utiles pour étudier le signal du point de vue de la mélodie, de l'harmonie, de la tonalité. L'étude de l'harmonie nécessite l'utilisation de fenêtres d'analyse plus grandes que pour le timbre qui sont typiquement de l'ordre d'une centaine de millisecondes. Par exemple, Müller calcule les chroma sur des trames de 200 ms espacées de 100 ms [MEK09].

Le contenu harmonique d'une portion de signal musical est généralement décrit par l'intermédiaire d'un vecteur de chroma. Il est habituellement défini comme un vecteur de dimension 12 qui représente l'énergie moyenne associée à chaque demi-ton de la gamme chromatique, sur l'ensemble des octaves audibles par une oreille humaine

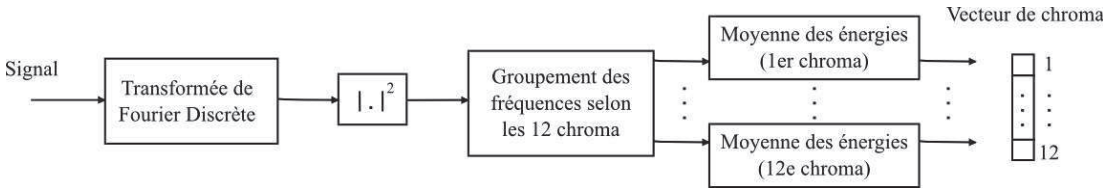

FIGURE 2.3 – Processus de génération des vecteurs de chroma.

[BW05]. Il s'appuie sur les travaux de Shepard [She64] qui propose de décrire une hauteur perçue suivant l'octave dans lequel elle se situe (*tone height*), et par la position de cette hauteur relativement à l'octave au sein duquel elle se situe (*chroma*). Bartsch et Wakefield ont introduit ce descripteur pour l'estimation de structure [BW01]. Pour l'obtenir, on calcule le spectre en puissance de la portion de signal considérée. Chaque fréquence du spectre est étiquetée selon le demi-ton qui lui est le plus proche. On calcule ensuite la moyenne arithmétique des énergies des fréquences associées à chaque demi-ton sans distinction d'octave pour obtenir chaque coefficient du vecteur de chroma. Ce procédé est illustré dans la figure 2.3.

Quelques autres versions du vecteur de chroma ont depuis été proposées. Notons toutefois que certains travaux considèrent ce vecteur à des résolutions fréquentielles différentes [Góm06] ou le calculent pour plusieurs bandes de fréquences [PK06]. Goto propose de remplacer l'étape de groupement par un filtrage du spectre en puissance [Got03]. D'autres variantes ont été proposées afin d'améliorer leur robustesse vis-à-vis des fluctuations de timbre et d'intensité sonore au cours du temps [Góm06, MEK09].

Un nouveau descripteur d'harmonie a récemment été proposé pour l'estimation de structure : le *multi-probe histogram* est un vecteur regroupant les probabilités de transition entre les chroma dominants de deux fenêtres d'analyses successives [KS10]. Il est fondé sur l'hypothèse que la structure tonale d'un morceau de musique est construite à partir d'un nombre limité de notes et d'intervalles musicaux.

Rythme Quelques approches considèrent des descripteurs relatifs au contenu rythmique des signaux étudiés [Jen07, PK08a]. Il s'agit ici d'une part de caractériser localement le signal musical en terme d'émission de notes (attaques) et de sons percussifs, et d'autre part de repérer les occurrences multiples des motifs issus de cette caractérisation. De tels descripteurs se basent sur l'analyse de l'évolution en terme d'intensité sonore et de spectre. C'est le cas du *rythmogramme* [Jen07], qui est obtenu en calculant l'autocorrélation du flux spectral perceptif (ou *perceptual spectral flux*). Ce flux est une fonction qui quantifie à chaque instant la puissance liée à l'ensemble des fréquences perçues dans son voisinage. Ce descripteur permet ainsi d'obtenir à la fois des informations sur le rythme et son évolution au cours du morceau.

Utilisation de l'échelle des temps comme échelle d'analyse du morceau La plupart des approches faisant intervenir une analyse structurelle se fondent sur une description du contenu musical exprimée à l'échelle des temps musicaux [Jeh05], [MND09], [PK06]. Ceci permet de disposer d'une représentation invariante par rapport à la vitesse d'exécution du morceau, qui constitue un bon compromis entre la finesse de la description et le temps de traitement.

En pratique, les vecteurs de descripteurs sont tout d'abord calculés pour des trames

de durée fixée se recouvrant de moitié. On associe ensuite à chaque temps musical la moyenne [WB10] ou la médiane [MND09] des vecteurs de descripteurs appartenant à un voisinage du temps considéré, et dont la durée est typiquement égale à celle d'un temps. Les instants des temps sont estimés à l'aide d'algorithmes développés en MIR (voir par exemple [Dav07], [EK10]). Les trames sont alors généralement de l'ordre de 500 ms.

Approche multi-strate Un nombre grandissant de travaux utilisent conjointement des descripteurs de timbre et d'harmonie [KS10, PK06, LNS07, CL11a]. Quelques-uns utilisent de plus des descripteurs de rythme [PK08a, Jen07].

Plusieurs descripteurs *dynamiques* de timbre et d'harmonie ont été proposés pour l'analyse structurelle [PLR02, BCL10, PKA11]. Ceux-ci prennent en compte l'évolution locale des descripteurs associés (MFCC, chroma) et leur ajoutent de ce fait une information de type rythmique.

2.4 Critères audio

Nous avons vu quelles strates musicales étaient habituellement considérées pour effectuer l'estimation de la structure, et par l'intermédiaire de quels descripteurs. Nous allons maintenant nous attacher à la manière dont ces descripteurs sont utilisés pour caractériser d'organisation du contenu des segments structurels.

Les méthodes actuelles ont principalement recours à deux familles de *critères audio* : l'homogénéité et la répétition [4]. Ces deux familles de critères peuvent-être qualifiées d'"externes aux segments", puisqu'ils comparent le contenu musical du segment à son environnement immédiat dans le cas de l'homogénéité, ou au reste du morceau dans le cas de la répétition. Ces critères peuvent être formulés comme des indices de présence des segments structurels (prenant des valeurs élevées aux instants associés à un segment structurel), ou comme des indices de présence de frontières structurelles (prenant des valeurs élevées aux instants de "rupture" entre segments).

2.4.1 Homogénéité

Définition Selon Pierre Schaeffer, un son est homogène si "son timbre, sa tessiture [5] et sa dynamique sont constants pendant toute sa durée" ([Sch52], p. 226). Plus précisément, la perception d'une telle "constance" résulte de la stabilité statistique de ces propriétés. Un son sinusoïdal peut être perçu comme "constant", bien que sa forme d'onde n'est pas constante à une échelle très fine. En s'inspirant de cette observation pour une échelle plus grande, on peut caractériser l'homogénéité d'une portion d'un morceau de musique par la stabilité de ces trois propriétés musicales au cours du temps. Un grand nombre de travaux ont utilisé cette hypothèse afin de caractériser les segments structurels. En conséquence, les frontières structurelles sont caractérisées par des ruptures de stabilité.

4. Dans son état de l'art [PMK10], Paulus proposer de classer les différentes méthodes suivant trois approches : l'homogénéité, la "nouveauté" et la répétition. Les concepts liés à l'homogénéité et à la nouveauté nous semblent très liés, car la détection d'une nouveauté au cours du morceau au sens de Paulus correspond à la détection d'une rupture d'homogénéité. Nous avons ainsi choisi de regrouper ces deux approches sous le terme d'homogénéité.

5. La tessiture est l'"étendue moyenne des hauteurs que peut jouer un instrumentiste ou un chanteur" selon [Sir09].

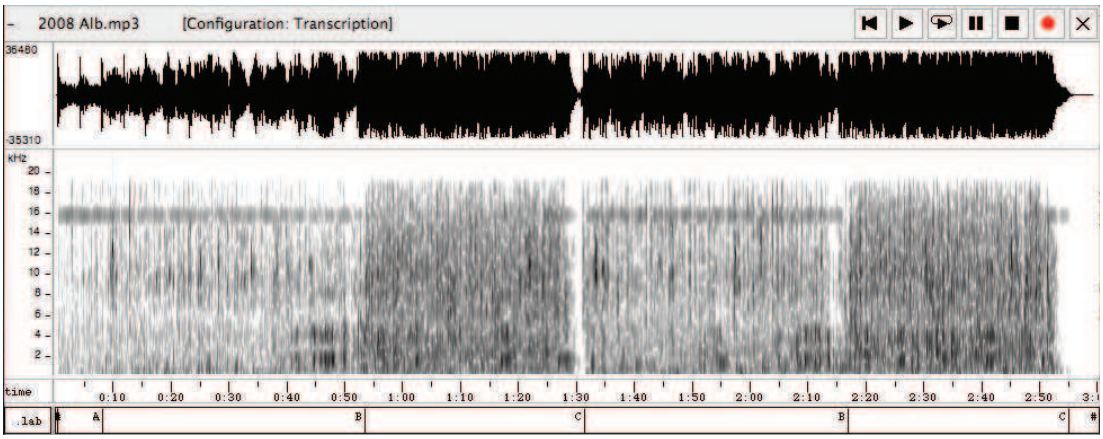

FIGURE 2.4 – Représentations temporelles de *Zemrën lamë peng* de Olta Boka. Sont représentés de haut en bas : la forme d'onde du morceau, son spectrogramme, puis l'annotation structurelle en couplets/refrains du chapitre 1.

Illustration On peut observer que dans un grand nombre de cas, l'instrumentation et la manière dont les instruments sont joués sont relativement stables sur un ensemble de portions des morceaux de musique.

Reprenons l'exemple du morceau *Zemrën lamë peng* de Olta Boka mis en regard de l'annotation structurelle en couplets/refrains du chapitre précédent (*cf.* l'annotation 3 de la figure 1.3). Pour faciliter la lecture de cet exemple, on associe respectivement aux étiquettes *intro*, *couplet* et *refrain* les lettres **A**, **B** et **C** comme dans la figure 2.4. Les instruments intervenant dans le morceau sont le chant, une ou deux guitares, une basse et des percussions (batterie et effets électroniques). Comme évoqué au chapitre 1, il est possible de déterminer deux "régimes" stables du point de vue du timbre et de la dynamique à l'écoute du morceau. Dans le premier, les instruments jouent doucement. La guitare présente possède un son clair, la basse et les percussions sont discrètes à l'écoute. La faible densité sonore qui en résulte est visible sur la forme d'onde et le spectrogramme entre 08 s et 53 s, puis entre et 1 min 32 et 2 min 17, ce qui correspond aux segments étiquetés **B**. Le second régime correspond à une densité sonore plus forte : on peut y entendre deux guitares saturées, la basse est plus présente et la batterie est jouée plus fort. Il se démarque ainsi du régime précédent comme on peut l'observer sur la forme d'onde et le spectrogramme entre les instants 53 s et 1 min 32, puis 2 min 17 et 2 min 55 (segments étiquetés **C**). Le segment d'introduction **A** est une version instrumentalement allégée des segments **B** : le chant y est absent. On peut lui associer un troisième régime, bien qu'il soit assez proche de celui associé aux **B**.

On remarque ainsi dans cet exemple que les différents régimes coïncident avec les frontières structurelles issues de l'annotation en couplets/refrains, d'où l'intérêt de prendre en compte un critère d'homogénéité pour l'analyse de la structure.

Descripteurs utilisés En pratique, les critères d'homogénéité sont calculés sur des descripteurs numériques relatifs au timbre [FC03, LS08], cependant certains travaux utilisent aussi les vecteurs de chroma [BCL10, LNS07]. L'observation de portions de morceau dont le contenu tonal est stable est interprétée en terme de tonalité.

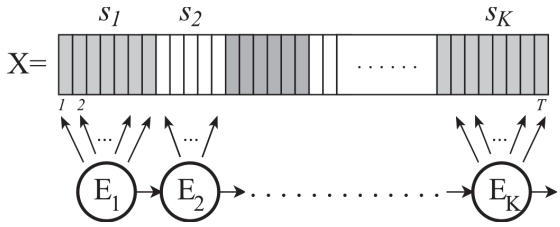

FIGURE 2.5 – Automate à états associé à la séquence de descripteurs d'un morceau de musique.

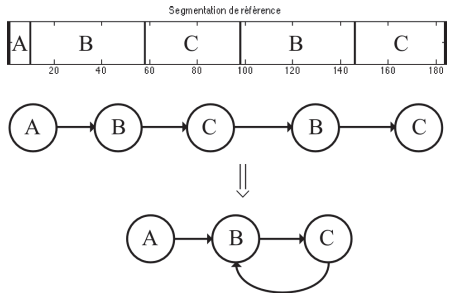

FIGURE 2.6 – Automate à états correspondant au morceau *Zemrën lamë peng*.

Modélisation d'un morceau de musique du point de vue de l'homogénéité D'un point de vue théorique, chaque segment structurel peut être modélisé par un processus stationnaire à partir duquel est générée sa séquence de descripteurs. Les segments d'une même classe sont associés au même processus. Chaque processus correspond à un état d'un automate qui modélise l'ensemble du morceau.

Soit un morceau de musique décrit par une séquence de T descripteurs, composé de K segments structurels $\{s_k\}$ groupés en M classes. Si l'on note E_k l'état associé au segment s_k, on obtient le schéma de la figure 2.5. Cette modélisation correspond à la notion d'*approche par états* proposée par Peeters dans [PLR02].

La figure 2.6 représente l'automate que l'on peut obtenir en modélisant chaque régime stable perçu lors de l'écoute du morceau par un processus stationnaire. Il y a dans ce cas M=3 états différents qui produisent les vecteurs de descripteurs décrivant le morceau.

Visualisation des segments homogènes par les matrices de similarité L'homogénéité des segments structurels peut se visualiser à l'aide de la matrice de similarité Σ, qui décrit la similarité entre les trames du morceau complet.

Soit une séquence de vecteurs de descripteurs $x = \{x_n\}_{1 \leq n \leq T}$, avec $T \in \mathbb{N}$ (pour tout n, x_n est de dimension $d \in \mathbb{N}$), et σ une mesure de similarité entre deux vecteurs, on a $\Sigma = \sigma(x_i, x_j)_{1 \leq i,j \leq T}$.

Dans la littérature, σ dérive souvent du produit scalaire [Foo00, PK06] :

$$\sigma(x_i, x_j) = 0.5 + 0.5 \frac{< x_i | x_j >}{||x_i|| ||x_j||} \tag{2.2}$$

avec $< .|. >$ le produit scalaire et $||.||$ la norme euclidienne.

Mais d'autres mesures peuvent être utilisées, dérivées de la distance euclidienne [Jen07] ou de la corrélation de Pearson [MND09]. On obtient ainsi une valeur maxi-

male lorsque deux vecteurs sont semblables, une valeur proche de 0 lorsqu'ils sont très différents.

Cette matrice est ensuite visualisée comme une image, en effectuant une association entre les valeurs de distance et une échelle de couleurs.

La figure 2.7 représente la matrice de similarité calculée sur la séquence de vecteurs de descripteurs MFCC de dimension 20 (le coefficient d'ordre 0 compris) décrivant le morceau *Zemrën lamë peng*. Les vecteurs de descripteurs sont exprimés à l'échelle des temps musicaux issus de l'estimateur de Ellis[6] puis sous-échantillonnés d'un facteur deux afin d'améliorer ici la lisibilité de la figure. La matrice est visualisée en niveaux de gris, les pixels les plus sombres correspondent à une forte similarité entre les vecteurs de descripteurs correspondants, et mise en regard des frontières structurelles de référence précédemment utilisées. On observe un ensemble de zones carrées de textures spécifiques alignés sur la diagonale de la matrice. On peut facilement faire correspondre les instants associés aux frontières entre ces zones avec les frontières de l'annotation structurelle en couplets/refrains. On remarque que les segments associés à la classe **C** correspondent à des textures très semblables en comparaison des segments de la classe **B**. Les MFCCs renseignant sur le timbre, ils mettent en valeur le fait que les segments **C** sont interprétés de la même manière, tandis que celle des segments **B** changent notamment par les percussions. Notons que malgré une différence de contraste, la structure des textures associés à ces derniers est semblable.

Méthodes d'analyse structurelle fondées sur l'hypothèse d'homogénéité des segments structurels Un certain nombre de méthodes utilisent une caractérisation des segments en terme d'homogénéité.

Plusieurs méthodes dérivent un critère d'homogénéité d'une matrice de similarité afin d'estimer la position des frontières structurelles. Foote a proposé une courbe de nouveauté timbrale qui résulte de la corrélation le long de la diagonale de cette matrice issue des MFCCs avec un noyau en damier [Foo00]. Cette courbe associe à chaque trame une valeur élevée lorsque elle est située à la frontière entre deux zones de texture distinctes, et une valeur minimale lorsqu'elle est contenue dans une zone de texture homogène. La position des frontières structurelles est ainsi estimée par la sélection de ses maxima les plus significatifs, par un seuillage direct [Foo00] ou un seuillage adaptatif [KS10, PLR02]. Cette méthode étant très utilisée, elle est décrite plus en détail en 2.6. Jensen évalue au contraire l'homogénéité locale le long de la diagonale avec un noyau constant obtenant des valeurs maximales pour les trames appartenant à des textures homogènes [Jen07]. Il segmente ensuite le morceau courant en sélectionnant le chemin de plus bas coût d'un graphe acyclique orienté dont chaque noeud représente un segment structurel possible. Chaque arc de ce graphe est pondéré par un coût combinant la mesure de l'homogénéité du segment décrite ci-dessus et une pénalité fixe associée à la sélection d'un nouveau noeud.

D'autres méthodes dérivent un critère d'homogénéité à partir d'une modélisation "par états" du morceau de musique [PLR02]. Après avoir découpé le signal audio soit en trames de durée fixée soit selon les temps musicaux, il s'agit de les regrouper en classes d'équivalences suivant leur contenu musical représenté par leurs descripteurs respectifs. Ces classes peuvent directement correspondre aux étiquettes structurelles [LC00], ou à des étiquettes d'un niveau intermédiaire [LS08]. Plusieurs travaux uti-

6. http ://labrosa.ee.columbia.edu/projects/coversongs/

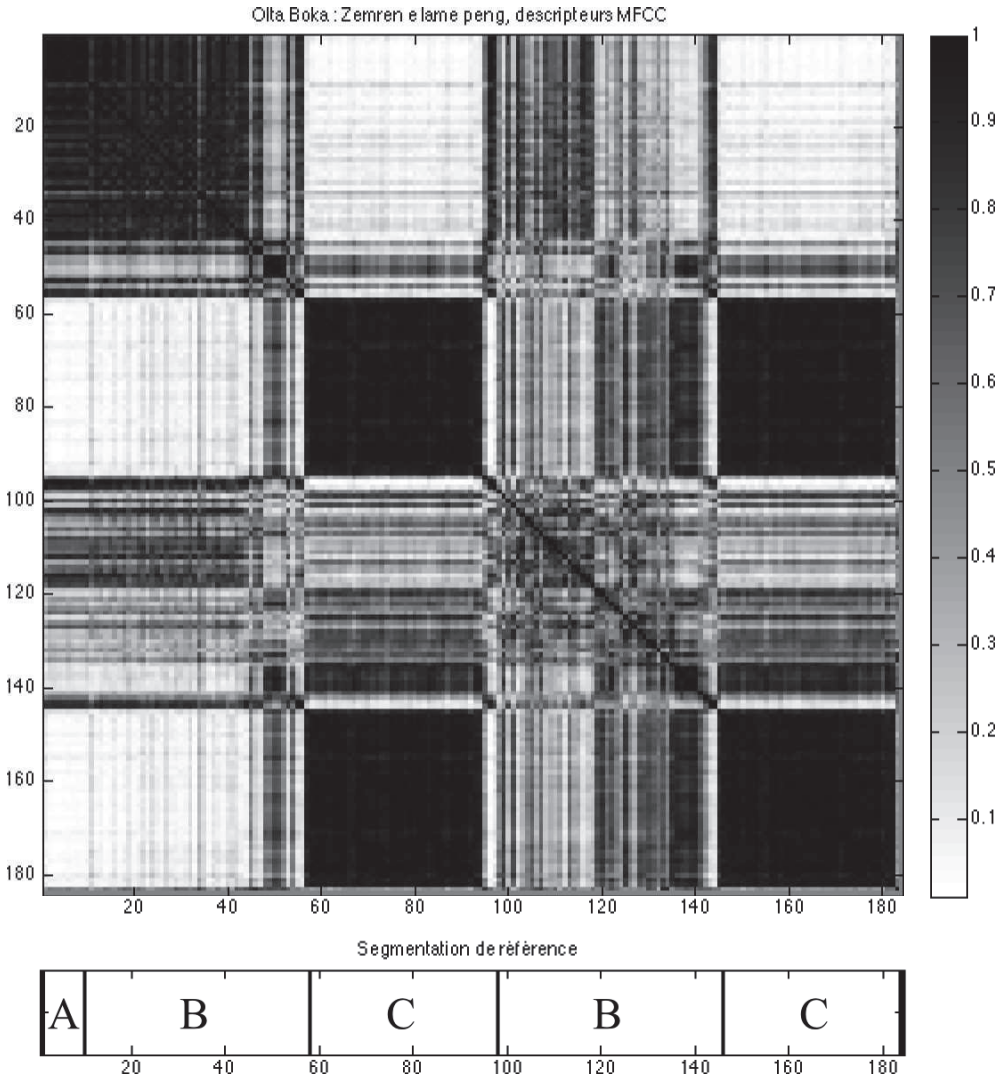

FIGURE 2.7 – Matrice de similarité de *Zemrën lamë peng* calculé sur la séquence de vecteurs contenant les 20 premiers coefficients MFCC (haut). L'annotation structurelle de référence correspond à la structure couplets/refrains du chapitre 1.

lisent des techniques de regroupement, ou *clustering*, dans ce but. Logan et Chu effectuent une pré-segmentation de la séquence de descripteurs en portions de taille 1 s, modélisent chaque portion par une Gaussienne et effectuent un *clustering* hiérarchique en associant chaque *cluster* à une classe de segments structurels [LC00]. Barrington *et al.* dérivent tout d'abord de la séquence de descripteurs du morceau un ensemble de modèles probabilistes (mélange de texture dynamiques) représentant les classes de segments à l'aide d'un algorithme espérance-maximisation, puis réalisent un clustering des trames sous contrainte par l'intermédiaire d'un algorithme des modes conditionnels itérés [BCL10]. D'autres méthodes de clustering ont recours à des techniques de recuit simulé [ANS+05, RCAS06] afin d'estimer la structure des morceaux de musique dans un cadre bayésien. D'autres travaux utilisent les modèles de Markov cachés (MMC), en associant la séquence de descripteurs à des observations, et les états cachés aux classes de segments [LC00, AS01, LS08]. Ces états sont modélisés par des distributions de probabilités gaussiennes dont sont issues les probabilités d'observation du modèle. Les paramètres des distributions des états et les probabilités de transition entre états sont appris de la séquence de descripteurs via un algorithme de type Baum-Welch, et la meilleure séquence d'états cachés expliquant les observations est obtenue par un algorithme de Viterbi.

Certains travaux ont combiné ces approches. Foote et Cooper [FC03] proposent de regrouper les segments obtenus par le seuillage de la courbe de nouveauté en classes par un *clustering* basé sur une décomposition en valeurs singulières reposant elle aussi sur une modélisation gaussienne des descripteurs des segments. Dans [PLR02] le morceau est segmenté en seuillant un critère proche de celui de Foote, et la valeur moyenne des descripteurs associés aux segments obtenus est utilisée afin d'initialiser les modèles de classes de segments d'un *clustering* de type K-moyennes. Les clusters obtenus sont ensuite utilisés pour initialiser les probabilités d'observation d'un MMC qui donnera l'estimation finale de la structure du morceau.

Une description plus détaillée des MMC et de l'algorithme de Viterbi est proposée en 2.6.

2.4.2 Répétition

Notion de répétition Selon Middleton, "la répétition est une caractéristique de toute musique, et un haut niveau de répétition peut constituer une marque spécifique au genre populaire" [Mid90]. Le contenu musical des morceaux que l'on considère dans le cadre de cette thèse présente en général une forte redondance au cours du temps, y compris à l'échelle structurelle. Un grand nombre de travaux utilisent ainsi l'hypothèse selon laquelle la plupart des segments structurels sont caractérisés par leur répétition, exacte ou approximative, au cours du morceau. Ceci amène à considérer l'ordre temporel du contenu musical d'un segment, contrairement à l'homogénéité qui n'en dépend pas.

Descripteurs utilisés La plupart des méthodes de détection des répétitions se fondent sur une description du morceau en terme de contenu tonal. Les travaux de Bartsch et Wakefield ont mis en avant qu'une description du morceau par une séquence de vecteurs de chroma offrait une représentation plus adaptée à la recherche de motifs musicaux que si l'on utilisait des vecteurs MFCCs, sur une base de 90 morceaux de musique pop [BW01]. On peut interpréter cela par le fait qu'il est plus facile de percevoir l'organisation temporelle des hauteurs des motifs mélodiques ou des accords à l'écoute,

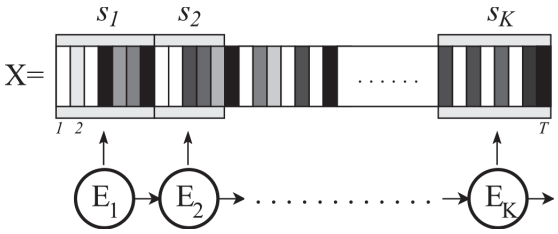

FIGURE 2.8 – Automate à états finis modélisant la séquence de descripteurs X d'un morceau de musique du point de vue de la répétition. Celui-ci est constitué de K segments structurels s_k, $k \in [1, K]$. Chaque état E_k est modélisé par un processus non-stationnaire, pour tout $k \in [1, K]$.

plutôt que de percevoir l'évolution du timbre, continue et plus "diffuse" au cours du morceau.

Modélisation d'un morceau de musique du point de vue de la répétition De la même manière que pour l'homogénéité, chaque segment structurel s_k est associé à un état E_k d'un automate modélisant le morceau. Chaque état peut être modélisé par un certain processus générant la séquence de vecteurs de descripteurs qui lui est associée. Les processus ne sont cependant pas stationnaires ici, afin de rendre compte de l'évolution temporelle de ces vecteurs. Les segments d'une même classe sont associés au même processus (*cf.* figure 2.8).

Visualisation des segments répétés par les matrices de similarité Les séquences de vecteurs de descripteurs répétés au cours du morceau sont localisables sur la matrice de similarité à l'aide des portions de sous-diagonales dont les coefficients sont associés à une forte similarité. L'observation de la matrice en niveaux de gris tel que le noir correspond à une similarité maximale permet d'observer des bandes diagonales sombres.

La matrice de similarité issue de la séquence de vecteurs de chroma de *Zemrën lamë peng*, exprimée à la même échelle que les MFCCs, est représentée dans la figure 2.9. Elle fait apparaître un ensemble de bandes sombres (mises en valeur par des lignes oranges) dont la détection des extrémités permet une estimation relativement précise des frontières d'une des annotations structurelles plausibles présentées dans le chapitre 1 (celle-ci est redonnée en bas de la matrice). Notons que la matrice ne permet pas de faire correspondre les segments de la première moitié du morceau (de 0 s à 1 min 31) avec ceux de la deuxième moitié (de 1 min 31 à 3 min 00). Ceci s'explique par le fait que cette première partie est répétée à une tonalité différente.

Cet exemple permet donc de mettre en valeur l'utilité d'un critère de répétition pour l'estimation de structure.

Méthodes d'analyse structurelle fondées sur l'hypothèse de répétition des segments structurels L'hypothèse de répétition des segments structurels est considérée pour l'estimation de leurs frontières ainsi que pour leur étiquetage. La majeure partie des méthodes s'appuient sur une analyse de la matrice de similarité, c'est-à-dire sur la recherche de séquences de coefficients matriciels correspondants à une forte similarité sur les sous-diagonales. D'autres approches traitent directement la séquence de

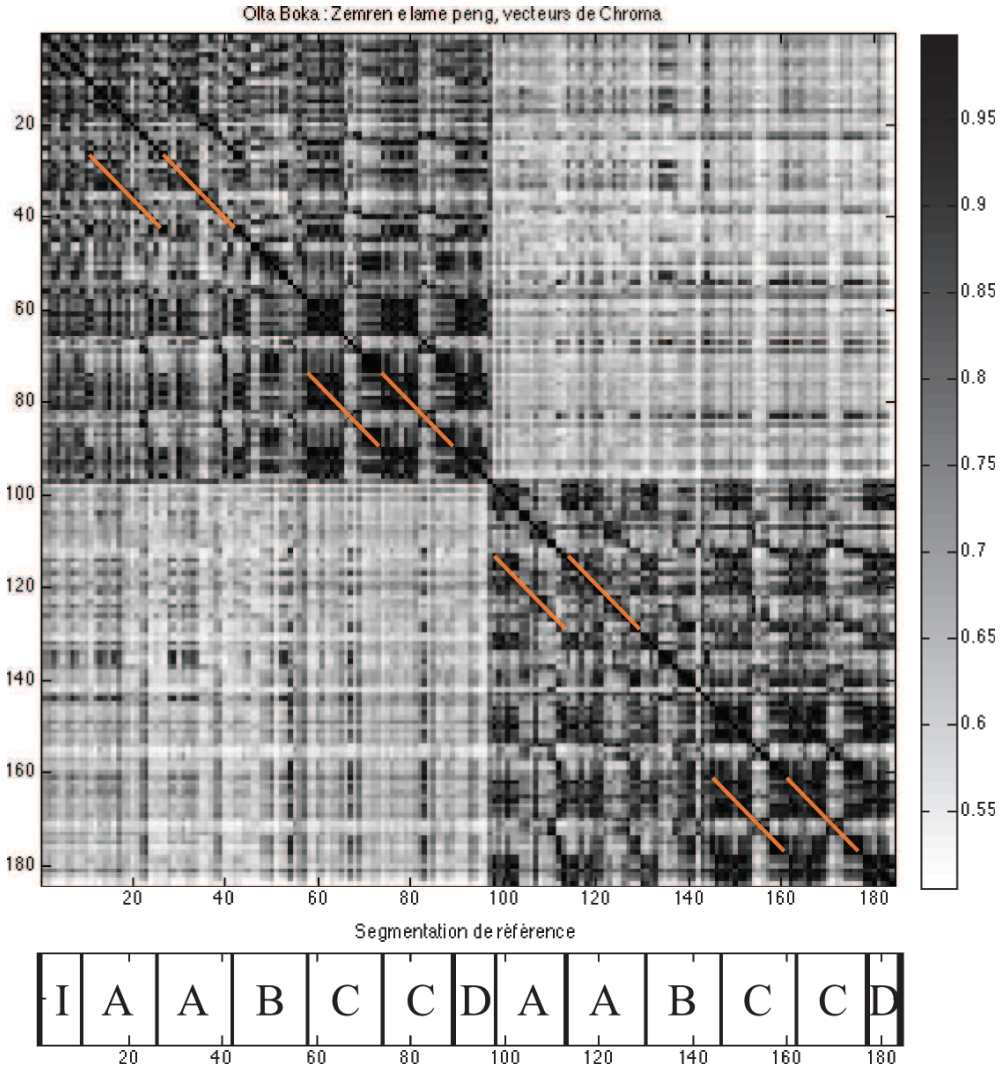

FIGURE 2.9 – Matrice de similarité de *Zemrën lamë peng* calculé sur les vecteurs de chroma (haut). L'annotation structurelle de référence (bas) correspond à l'annotation 1 présentée au chapitre 1. Les bandes diagonales sombres correspondant aux séquences de descripteurs répétées au cours du morceau sont mises en valeur par des lignes oranges dans la partie triangulaire inférieure de la matrice.

descripteurs afin de repérer les motifs répétés.

Cette matrice subit souvent un pré-traitement afin de mettre en valeur les portions de sous-diagonales de forte similarité, et d'améliorer la robustesse aux répétitions approximatives. Il peut s'agir d'un lissage par filtre moyen [BW01] ou filtre médian [MND09], d'opérations morphologiques d'érosion ou de dilatation [LWZ04] issues du traitement de l'image, ou de l'utilisation de matrices de similarité d'ordre supérieur [Pee07].

Les séquences répétées sont ensuite localisées par l'intermédiaire de portions de sous-diagonales de forte similarité. Le début des séquences "significatives" c'est-à-dire le plus souvent répétées ou les plus longues, est détecté par seuillage de la matrice. Chacun d'entre eux constitue le début d'un chemin de coefficients de forte similarité qui est "propagé" en diagonale jusqu'à ce que sa similarité moyenne passe sous un certain seuil. On peut y parvenir par filtrage [Got03, MND09] ou par des méthodes de programmation dynamique (DTW pour *Dynamic Time Warping*) [DH02].

Une fois les segments répétés localisés, ils sont regroupés en classes selon leur taux de recouvrement ou le "bon alignement" de leurs séquences de descripteurs respectifs [DH02, MND09, Got03], en veillant à ce que les segments d'une même classe soient de taille comparable et ne se recouvrent pas temporellement [Pee07, PK06]. Ce regroupement s'effectue généralement par comparaison des vecteurs de chroma, mais d'autres espaces ont été utilisés. Kaiser et Sikora décrivent par exemple chaque segment par ses coefficients sur la diagonale de la matrice de similarité du morceau, projetés sur une base de vecteurs obtenus par factorisation matricielle positive, ou NMF pour *Non-negative Matrix Factorization* [KS10]. Weiss et Bello [WB10] identifient un ensemble de motifs harmoniques répétés sur la séquence de vecteurs de chroma qui décrit le morceau de musique. Pour ce faire, ils ont recours à l'algorithme *Shift-Invariant Probabilistic Latent Component Analysis* (SI-PLCA), version probabiliste de la NMF, qui intègre des contraintes de parcimonie afin d'estimer le nombre de motifs et leur taille. Chaque motif représente une classe de segment, et ils associent par la suite à chaque trame le motif qui lui est le plus proche, afin d'estimer la structure entière.

Shiu *et al.* [SJJ06] combinent la détection des répétitions et leur regroupement. Ils interprètent la matrice de similarité comme une matrice de transition entre états. Par l'intermédiaire d'une fonction de coût qui privilégie les déplacements diagonaux sur cette matrice, ils recherchent le chemin le plus probable, c'est-à-dire de plus forte similarité, parcourant la partie triangulaire supérieure (ou inférieure) en respectant l'ordre temporel. Chaque répétition détectée renvoie à une paire de segments similaires du point de vue de leur évolution temporelle. Les classes de segments sont principalement formées en regroupant les segments dont les séquences de descripteurs s'alignent approximativement [Got03, DH02].

Notons enfin que d'autres approches se fondent sur l'alignement de séquences à l'aide de modèles de Markov cachés [RC07, AS02].

2.4.3 Utilisation de plusieurs critères ou plusieurs strates pour l'estimation de structure

Les méthodes n'utilisent en général qu'un seul critère pour effectuer la segmentation et l'étiquetage, typiquement l'homogénéité pour la segmentation et la répétition pour l'étiquetage [PK06, KS10]. La combinaison de ces deux critères a commencé à être explorée pour l'étiquetage de segments déjà localisés par un critère d'homogénéité

dans [PK08a]. Dans ce cadre, la combinaison des critères, exprimés sous la forme de probabilités, est effectuée à l'aide d'une moyenne géométrique.

Un des axes de recherche de cette thèse, développé dans la partie 4.1, porte sur la combinaison de critères pour la segmentation.

Mentionnons les récents travaux de Chen et Li [CL11a] qui utilisent une approche multi-strate et mono-critère. Ici, les auteurs utilisent le même processus sur les descripteurs d'harmonie et de timbre afin d'obtenir deux estimations de la structure du morceau, à savoir un clustering hiérarchique à deux niveaux. Ces estimations sont utilisées afin de calculer une matrice de similarité lissée et de basse résolution segmentée à l'aide d'une approche par NMF.

2.5 Contraintes structurelles

L'analyse du contenu musical par l'intermédiaire des critères permet d'aboutir à un ensemble de structures possibles pour un même morceau de musique. Pour pouvoir converger vers une structure unique, il est nécessaire d'orienter la recherche par l'intermédiaire d'hypothèses sur la structure musicale, que l'on nommera par la suite des contraintes structurelles. À notre connaissance, cette composante du problème d'estimation de structure a été peu étudiée jusqu'à présent.

Les contraintes structurelles de l'état de l'art sont de quatre ordres :
- sur le nombre de segments structurels,
- sur leur durée,
- sur le nombre de classes de segments (étiquettes structurelles),
- sur le nombre de segments par classe.

Notons que ces contraintes ne sont pas indépendantes : décrire un morceau avec un grand nombre de segments structurels implique qu'ils soient de courte durée, et la diminution du nombre d'étiquettes structurelles augmente le nombre de segments par classe.

Jensen associe à un morceau de musique la segmentation qui va minimiser un certain coût de segmentation [Jen07]. En plus de prendre en compte l'homogénéité du contenu timbral des segments structurels, ce coût comprend une quantité positive proportionnelle au nombre de segments afin de privilégier les segmentations constituées d'un petit nombre de segments.

Plusieurs travaux font intervenir des hypothèses sur la durée des segments. Plusieurs modèles probabilistes de leur durée ont été proposés [LS06, ANS+05]. Mauch *et al.* considèrent que la durée d'un segment correspond à un nombre entier de mesures musicales [MND09]. [PK06, Pee07] imposent que les segments associés à une même étiquette soient de durée comparable.

Paulus et Klapuri réalisent un groupement des segments détectés par la minimisation d'un coût qui pénalise les structures constituées d'un grand nombre d'étiquettes et celles qui contiennent des étiquettes associées à un seul segment [PK06]. Weiss et Bello limitent le nombre de classes de segments par l'intermédiaire de distributions de probabilité parcimonieuses dans [WB10].

Dans le cadre de la thèse, nous nous intéressons à une contrainte particulière sur la durée des segments structurels. Cette contrainte est motivée par une hypothèse de pulsation structurelle, qui considère que le nombre de temps de la plupart des segments structurels est comparable d'un segment à l'autre, sans distinction de classe. Cette nouvelle hypothèse est introduite dans la partie 3.4.3.

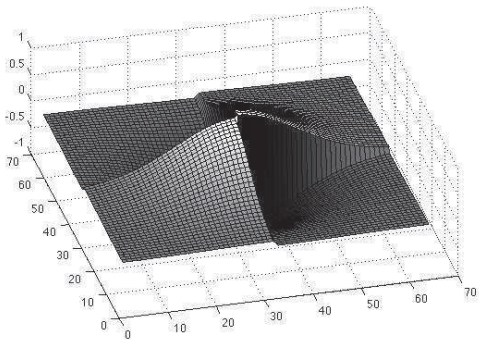

FIGURE 2.10 – Noyau de corrélation en damier de taille 64 par 64 lissé à l'aide d'une fenêtre gaussienne

2.6 Outils

Comme nous l'avons vu dans les parties 2.4.1 et 2.4.2, les différentes méthodes d'analyse structurelle ont recours à une variété d'outils, principalement issus du traitement de l'image et des statistiques. Nous donnons ici quelques détails concernant les approches importantes évoquées dans ce chapitre qui sont en lien avec les contributions algorithmiques que nous présentons dans le chapitre 4.

Fonction de nouveauté Nous avons vu dans la partie 2.4.1 que la matrice de similarité permet de visualiser les portions de signal homogènes d'un morceau de musique par l'intermédiaire de zones carrées de textures spécifiques le long de la diagonale de sa matrice de similarité. La fonction de nouveauté est un outil utile pour localiser les frontières de telles zones, qui correspondent à des instants de rupture d'homogénéité.

Elle est obtenue en effectuant une corrélation entre la matrice de similarité issue des coefficients MFCC extraits du morceau de musique avec un noyau de corrélation particulier. Celui en damier, lissé à l'aide d'une fenêtre gaussienne, est généralement utilisé [Foo00, FC03, KS10]. Un tel noyau noté C et de taille L, est défini comme suit :

$$C(i,j) = \begin{cases} 1 & \text{si} \quad 1 \leq i,j \leq \frac{1}{2} \quad \text{ou} \quad \frac{1}{2} \leq i,j \leq L \\ -1 & \text{sinon} \end{cases} \tag{2.3}$$

Par exemple, si $L = 4$, on obtient :

$$C = \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{pmatrix} \tag{2.4}$$

En pratique, le noyau est lissé afin d'éviter les effets de bord, par exemple à l'aide d'une fenêtre gaussienne comme montré dans la figure 2.10.

Soient $S = \{s_{i,j}\}_{1 \leq i,j \leq N}$ la matrice de similarité issue du morceau courant, et $C = \{C_{i,j}\}_{1 \leq i,j \leq L}$ le noyau de corrélation en damier, avec N et $L \in \mathbb{N}$, la fonction de nouveauté F se calcule comme suit :

$$F(i) = \sum_{m=-L/2}^{L/2} \sum_{n=-L/2}^{L/2} C(m,n) S(i+m, i+n) \tag{2.5}$$

Cette fonction est détaillée dans [Foo00].

Modèle de Markov Caché (MMC) Le modèle de Markov caché (ou *Hidden Markov Model*) est un modèle statistique fréquemment utilisé dans les domaines du traitement de la parole et de l'audio pour la reconnaissance de motifs. Il permet de modéliser une séquence d'observations $x = \{x_n\}_{1 \leq n \leq T}$ ($T \in \mathbb{N}$) par une séquence d'états *cachés* $q = \{q_n\}_{1 \leq n \leq T}$ nondirectement observables.

Un MMC est paramétré par :

- un ensemble d'états $E = \{E_m\}_{1 \leq m \leq M}$, avec $M \in \mathbb{N}$,
- un ensemble de probabilités de transition entre états $a_{i,j} = p(q_t = E_i | q_{t-1} = E_j)$, avec $1 \leq i, j \leq M$,
- un ensemble de probabilités initiales $\Pi_m = p(q_1 = E_m)$, avec $1 \leq m \leq M$,
- un ensemble de probabilités d'observation $b_t(m) = p(x_t | q_t = E_m)$, avec $1 \leq m \leq M$.

Une description détaillée du MMC et des principaux algorithmes associés, comme l'algorithme de Baum-Welch pour l'apprentissage de ses paramètres, est proposée dans [Rab89].

Algorithme de Viterbi L'algorithme de Viterbi permet d'estimer la séquence d'états cachés $q = \{q_t\}_{1 \leq t \leq T}$ la plus probable correspondant à une séquence d'observations $x = \{x_t\}_{1 \leq t \leq T}$ modélisée par un MMC. Supposons que les paramètres $\Theta = \{A, B, \Pi\}$ du MMC sont connus, avec $A = \{a_{i,j}\}_{1 \leq i,j \leq M}$, $B = \{b_t(m)\}_{1 \leq m \leq M, 1 \leq t \leq T}$ et $\Pi = \{\Pi_m\}_{1 \leq m \leq M}$.

Soit $\delta_t(i)$ la probabilité associée à la séquence d'états se terminant par l'état E_i expliquant le mieux la séquence d'observations $\{x_1, x_2, ..., x_t\}$:

$$\delta_t(i) = \max_{q_1, q_2, ..., q_{t-1}} p(q_1, q_2, ..., q_t = E_i, x_1, x_2, ..., x_t | \Theta) \tag{2.6}$$

On a, récursivement :

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{i,j}] b_{t+1}(j) \tag{2.7}$$

On conserve l'argument qui maximise la formule (2.7) pour tout t et j dans $\Delta_t(j)$. Une procédure de *backtracking* partant de l'état final et sélectionnant récursivement les arguments optimaux stockés dans $\{\Delta_t(j)\}$ jusqu'à l'état initial permet d'obtenir la séquence d'états la plus probable correspondant à x. Cette procédure est détaillée dans [Rab89].

Déformation temporelle dynamique La déformation temporelle dynamique (ou DTW de l'anglais *Dynamic Time Warping*) est une technique d'alignement de séquences. Dans le cadre de l'estimation de la structure d'un morceau de musique, elle est utile pour rechercher des séquences temporelles de descripteurs qui se répètent avec des variations de durée ou de vitesse.

Plus précisément, étant donnés s_1 et s_2 deux séquences de descripteurs, on considère la matrice de distance qui leur est associée. Plusieurs distances peuvent être utilisées : distance de Manhattan, distance euclidienne, *etc.* L'alignement des séquences correspond à la sélection des distances entre descripteurs qui forment un chemin partant de l'origine de la matrice (distance entre le premier descripteur de s_1 et de s_2) à la fin de

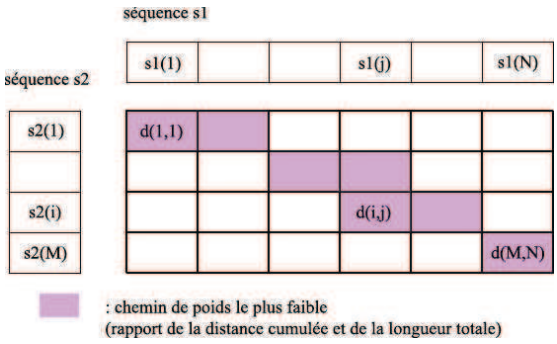

FIGURE 2.11 – Alignement de deux séquences de descripteurs s_1 et s_2 à l'aide de la matrice de distances associée.

la matrice (distance entre le dernier descripteur de s_1 et de s_2), comme le montre la figure 2.11.

Afin de conserver l'ordre des descripteurs dans les deux séquences, les chemins considérés sont connexes et contraints de suivre les directions horizontale de la gauche vers la droite (suppression de descripteurs de s_1 pour son bon alignement avec s_2), verticale de haut en bas (insertion de descripteurs dans s_1 pour son bon alignement avec s_2), et diagonale de la gauche vers la droite et de haut en bas (correspondance entre s_1 et s_2).

Le meilleur alignement entre s_1 et s_2 correspond au chemin qui satisfait ces contraintes et dont la distance totale est minimale. L'estimation de ce choix s'effectue par un algorithme dit de programmation dynamique semblable à celui de Viterbi (nous sommes dans un cas particulier des MMC). Pour plus de détails, voir [MRR80].

Autres outils Nous invitons le lecteur à la recherche d'informations sur les autres outils évoqués lors de cet état de l'art à se référer à plusieurs ouvrages. Un grand nombre de techniques de classification non-supervisée, comme le *clustering* hiérarchique et les K-moyennes sont décrites dans [AHdS96]. Les différentes techniques de factorisation matricielle évoquées comme la décomposition en valeurs singulières, la NMF et la SI-PLCA sont respectivement décrites dans [Wei99], [LPAK09] et [SRS08]. Enfin, [NWL+07] détaille les réseaux bayésiens.

2.7 Résumé du chapitre

Ce chapitre nous a tout d'abord permis de présenter un bref panorama des principaux travaux relatifs à la spécification de la structure musicale. Du point de vue musicologique, la majeure partie de ces approches sont spécifiques à un genre musical donné et sont fondées sur l'étude d'un nombre limité de strates musicales. Elles sont donc difficilement exploitables dans l'optique d'étudier les morceaux de musique "conventionnels". Du point de vue du MIR, il n'y a actuellement pas de spécification unanimement adoptée de la structure d'un morceau de musique. Nombre de travaux relevant de l'*analyse structurelle* proposent de simplifier ce problème en se concentrant sur un aspect particulier de la structure (détection de motifs répétés) ou sur une application particulière (extraction de vignettes sonores, résumé automatique).

Un état de l'art des méthodes d'analyse structurelle est ensuite proposé selon une grille de lecture personnelle constituée de trois composantes : les strates musicales considérées, les caractérisations des segments structurels sur ces strates (*critères audio*), et les hypothèses structurelles permettant de restreindre le nombre de structures visées (*contraintes structurelles*). Le caractère multidimensionnel de la musique est en général traité en considérant plusieurs strates musicales (timbre, harmonie, mélodie) au cours de l'analyse structurelle. Cependant, les segments structurels peuvent se caractériser de différentes manières selon les strates, et en pratique un seul critère est utilisé pour estimer leurs frontières dans le morceau. On note par ailleurs que les critères de l'état de l'art, l'homogénéité et la répétition, sont des critères de type externes aux segments et n'utilisent aucune information relative à leur structure interne. Le caractère multi-échelles de la musique est en général traité de manière implicite, par l'intermédiaire de contraintes structurelles sur le nombre de segments structurels du morceau, leur durée, le nombre d'étiquettes structurelles recherchées et le nombre de segments possédant la même étiquette.

Cet état de l'art permet de mettre en lumière la nécessité d'une spécification de la structure recherchée, ainsi que plusieurs pistes de recherche prometteuses pour son estimation : l'utilisation conjointe de plusieurs critères pour l'estimation des frontières structurelles, l'élaboration d'une caractérisation intrinsèque des segments structurels et d'une nouvelle contrainte structurelle.

Chapitre 3

Spécification et méthodologie d'annotation de la structure sémiotique

Le chapitre 1 nous a permis de présenter le problème méthodologique auquel nous sommes confrontés : il n'existe pas de caractérisation unique de la structure d'un morceau de musique. Nous avons vu que ce problème était notamment lié à trois aspects de la musique :

- la musique est un objet multi-strate : à quelle(s) strate(s) s'attacher pour l'étude de la structure ?
- l'organisation de la musique est multi-échelles : à quelle échelle d'étude travailler ?
- la notion de *similarité* de la musique est ambiguë : comment comparer des segments structurels et comment modéliser une classe de segments structurels ?

C'est dans cette optique qu'un groupe de travail s'est mis en place à l'IRISA, portant sur la spécification d'une définition opérationnelle de la structure musicale et sur l'élaboration d'une méthodologie d'annotation pour produire une description structurelle des morceaux de musique quasiment univoque et reproductible d'un annotateur à l'autre. L'approche qui en est issue est fondée sur l'expérience d'écoute de l'annotateur et non sur son expertise musicologique. Elle a pour but d'être applicable à une grande variété de styles et genres musicaux, et n'est pas spécifique à une application particulière. Ces travaux ont donné lieu à plusieurs publications sur l'annotation des frontières structurelles [BLSV10a, BLSV10b, BDSV11] et sur l'étiquetage des segments [BDSV12a, BDSV12c], et dont nous faisons la synthèse dans ce chapitre [1].

3.1 Pourquoi s'intéresser à une convention d'annotation de la structure ?

L'idée d'établir une convention sur la spécification et l'annotation de la structure fait actuellement débat. Devant la difficulté de la tâche, une approche statistique peut être intéressante : il s'agit de collecter un grand nombre d'annotations manuelles de la structure pour un ensemble de morceaux de préférence non limité à un genre musical particulier. Ceci permettrait d'une part l'étude des choix d'annotation les plus

1. La publication d'un manuel d'annotation est prévue dans les mois qui suivront cette thèse.

fréquents, d'autre part une évaluation cohérente des algorithmes développés, permettant leur diagnostic. Cependant la constitution d'une telle base est très coûteuse.

Nous soutenons ainsi que l'établissement d'une convention sur la spécification et l'annotation d'une structure musicale est nécessaire à condition qu'elles soient aussi génériques et reproductibles que possible. Une telle approche permettrait la production de bases d'annotations propice à l'élaboration et au diagnostic de systèmes d'estimation de structure.

3.2 Cadre d'étude

Le périmètre de notre étude couvre un éventail assez conséquent de genres musicaux, il nous paraît donc important que la notion de structure que l'on utilise soit aussi générique que possible. La linguistique, qui porte sur l'étude du fonctionnement des langues en général, s'est intéressée à un problème similaire pour le langage humain [Sch94]. Le *structuralisme* est un courant particulier de la linguistique, qui a été initié par Ferdinand de Saussure [dS16] et étendu ensuite à de nombreuses autres disciplines, notamment à la sémiologie musicale dans le cadre de la musique écrite [Ruw87, Nat87]. On peut en résumer ainsi l'axiome général : la structure d'une entité est essentiellement déterminée par les relations que ses constituants entretiennent les uns avec les autres au sein de l'entité, indépendamment de la forme et du sens de ces constituants.

Appliquée à notre objet d'étude, cet axiome nous amène à considérer un morceau de musique comme le résultat de l'agencement d'un ensemble d'éléments constitutifs, selon un certain processus *syntagmatique*. Les éléments constitutifs entretiennent également entre eux des relations *paradigmatiques* qui permettent de les comparer, et qui s'expriment sous la forme de *relations d'équivalence*. L'ensemble forme un *système* au sens structuraliste du terme, c'est-à-dire une "entité de dépendances internes", selon la définition de Hjelmslev [Hje43]. Le morceau de musique apparaît alors comme une réalisation particulière (ou *observation*) issue de ce système et le problème d'estimation de structure musicale consiste à déterminer, à partir de cette unique observation, la délimitation des éléments constitutifs du morceau (*segmentation* ou, plus généralement, *décomposition*) et l'attribution d'une classe d'équivalence à chacun d'entre eux (*étiquetage*).

On ajoute aux aspects syntagmatiques et paradigmatiques issus du structuralisme des considérations d'ordre morphologiques propres à la musique à l'aide d'un modèle d'organisation interne des segments structurels, qui est décrit par la suite.

3.3 La structure sémiotique

Comme nous l'avons évoqué au chapitre 1, l'organisation de la musique peut s'observer à plusieurs échelles. Nous nous proposons de nous pencher sur une structure macroscopique de type *sémiotique*, dont les éléments constitutifs ont une durée de l'ordre de 15 secondes. Le terme sémiotique est ici utilisé dans un registre assez restreint au regard des travaux de Nattiez [Nat87], afin de considérer une représentation symbolique et métaphorique de haut niveau du contenu musical. Ceci permet de se démarquer d'une représentation d'ordre *sémantique*, qui tend à donner une signification particulière à ses différents éléments et qui sort du cadre de notre étude.

La structure sémiotique d'un morceau peut ressembler à la séquence de symboles suivante :

ABCDEFBCDEGDEDEH

Elle permet de rendre compte :

- de la décomposition macroscopique de l'ensemble du morceau en un nombre limité d'éléments ou *blocs* de taille comparable, et
- d'un ensemble de relations d'équivalence ou de similarité entre les blocs matérialisées par leurs étiquettes. Notre exemple comprend ici 8 étiquettes distinctes, c'est-à-dire 8 classes de blocs.

La recherche d'une description sémiotique pour un morceau de musique donné nécessite tout d'abord d'identifier la *granularité* adéquate liée à la taille et au nombre de blocs, ce qui permet de conditionner l'inventaire des étiquettes structurelles. Dans le cadre de l'exemple précédent, le choix d'une granularité plus fine peut conduire à la séquence d'étiquettes

AA'BB'CC'DD'EE'FF'BB'CC'DD'EE'GG'DD'EE'DD'EE'HH'

pour laquelle chaque symbole X est systématiquement suivi par un symbole X', ce qui conduit à une description sémiotique très redondante et donc peu concise. À l'inverse, l'utilisation d'une granularité plus grossière peut impliquer le regroupement de blocs en unités de tailles diverses ou *irrégulières* qui sont plus hétérogènes, comme par exemple

A BC DE F BC DE G DE DE H

ou bien aboutir à une représentation beaucoup plus confuse

AB CD EF BC DE GD ED EH

qui ne permet pas de rendre compte des similarités existant entre les portions du morceau associées aux mêmes étiquettes à l'échelle inférieure.

Cet exemple permet d'illustrer le cas simple pour lequel il existe clairement une granularité permettant d'obtenir un compromis optimal entre :

- un alphabet d'étiquettes sémiotiques (symboles) de taille minimale,
- une séquence d'étiquettes informative, c'est-à-dire fournissant une description sémiotique aussi précise temporellement que complète du point de vue des propriétés musicales du morceau,
- des tailles de blocs comparables.

Cependant des cas plus complexes sont bien sûr rencontrés en pratique. Afin de traiter un grand nombre de cas, nous proposons un ensemble d'axiomes, de concepts et de principes méthodologiques afin d'identifier la granularité la plus appropriée pour décrire la structure sémiotique, de localiser de la manière la plus univoque possible des frontières de blocs correspondants et d'associer à chaque bloc une étiquette sémiotique.

3.4 Concepts et axiomes de travail

Notre approche se fonde sur plusieurs axiomes sur la nature des éléments structurels et sur leur organisation au sein du morceau.

3.4.1 Bloc structurel

Les *blocs structurels* désignent les éléments constitutifs de la structure sémiotique. Ils sont principalement agencés par concaténation, mais peuvent aussi se chevaucher (*tuilage*). Un bloc structurel est défini par un *début*, une *durée*, une *taille* et une *étiquette sémiotique*. La distinction entre durée et taille est explicitée dans la partie 3.4.2.

Un bloc peut se décomposer en un *radical*, c'est-à-dire une version *régulière* du bloc telle que définie dans la partie 3.4.4, et une ou plusieurs distorsions, telles que des troncatures ou des insertions.

On admet qu'un bloc est *autonome*, c'est-à-dire que son écoute donne l'impression d'une cohérence musicale propre. Cette notion est liée au caractère *cyclique* des processus ayant engendré le morceau de musique développé dans la partie 3.5.2, ainsi qu'aux relations particulières que partagent les éléments constitutifs du bloc lui-même, précisés dans la partie 3.5.3.

3.4.2 Taille de bloc

Il est nécessaire de choisir une échelle d'étude du contenu musical afin de comparer le contenu de différentes portions du morceau au cours du temps. L'échelle associée aux temps musicaux est intéressante car robuste aux variations de tempo, c'est-à-dire de la vitesse d'exécution du morceau. Cependant, si la période associée à ces temps est spécifiée pour la musique écrite, elle est ambiguë dans le cas d'un enregistrement sonore : typiquement, on pourra associer un morceau interprété à un tempo de 120 battements par minutes (bpm) les échelles de temps associées aux tempos 60 bpm et 120 bpm [MM04].

À l'échelle des temps musicaux, qui s'appuie sur des notions liées à la composition, on préférera une échelle de type perceptive, cohérente selon les annotateurs. On propose d'utiliser dans cette thèse l'échelle des *snaps*, en définissant le snap comme le multiple ou le sous-multiple du temps musical dont la période est la plus proche de 1 s (60 bpm) [2].

On distinguera ainsi la *durée* d'un bloc en secondes de sa *taille* en snaps. L'utilisation de cette taille permettra ainsi de faire correspondre deux blocs qui ne diffèrent que par leur tempo.

Cette définition du snap pourra être revue à l'avenir. Il serait utile d'ajouter une spécification supplémentaire permettant de rendre cette unité cohérente sur l'ensemble des morceaux de musique. On peut imaginer avoir à comparer des blocs structurels appartenant à plusieurs morceaux différents, par exemple dans le cadre d'un album-concept dans lequel plusieurs portions d'un morceau sont reprises dans d'autres morceaux du même album, ou plus généralement dans la détection de reprises ou *cover songs* en anglais.

3.4.3 Patron structurel et pulsation structurelle

Notre approche se fonde sur l'axiome suivant : la structure sémiotique peut être décrite en référence à un *patron structurel*, c'est-à-dire une partition prototypique de l'échelle des snaps ou des temps musicaux. Par exemple, un patron structurel très commun est la répétition de blocs de 16 snaps.

Si la structure haut-niveau d'un morceau de musique est gouvernée par un patron structurel, les blocs sémiotiques observés au cours du morceau résultent de la *réalisation* de ce patron, ce qui peut conduire à l'observation de blocs de taille irrégulière. Dans un grand nombre de cas les blocs irréguliers peuvent être ramenés à des *radicaux* réguliers qui se conforment au patron structurel, comme développé dans la partie 3.4.4.

2. Cette définition du snap se rapproche de celle du *tactus* sans pour autant l'équivaloir clairement. Cette nuance, même si elle devra être affinée à l'avenir, ne sera pas précisée dans la thèse car elle n'influence pas la définition de la structure sémiotique.

Les *blocs prototypiques* qui composent le patron structurel sont en général en nombre limité, ce qui implique l'existence d'un nombre limité de tailles prototypiques ou *pulsations structurelles* autour desquelles "gravitent" les tailles des blocs réalisés au cours du morceau. Par exemple, si la pulsation structurelle d'un morceau égale 16 snaps (c'est-à-dire que le patron structurel n'est composé que de blocs prototypiques de 16 snaps), la taille de certains blocs pourra être égale à 18. Notons que l'on peut trouver plusieurs pulsations structurelles différentes dans un morceau de musique.

3.4.4 Blocs réguliers, blocs irréguliers

On appelle *bloc régulier* un bloc observé au cours d'un morceau de musique dont la taille correspond à celle de son bloc prototypique, qui constitue un *radical structurel*.

Dans la majorité des cas, un bloc observé peut être associé à un bloc prototypique particulier sans que leurs tailles correspondent. Un tel bloc est dit *irrégulier*, et peut être interprété comme un radical structurel ayant subi une ou plusieurs distorsions d'ordre temporel.

S'il est plus court, il peut correspondre à un bloc radical auquel on a enlevé une ou plusieurs portions, au début, à la fin ou à l'intérieur de ce bloc. Par exemple, on peut observer la répétition de la première ou la seconde moitié d'un bloc régulier ailleurs au cours d'un morceau.

S'il est plus long, il est souvent possible de le réduire à un *radical* et un ou plusieurs *affixes*. Un affixe est une portion de bloc qui peut être vue comme une insertion dans le radical. Il peut consister en la reprise d'une portion du radical auquel il est rattaché ou ne partager aucun lien avec lui. Si l'insertion de l'affixe intervient en début de bloc (respectivement en fin de bloc), il est nommé *préfixe* (respectivement *suffixe*).

3.5 Principes méthodologiques pour l'annotation de la structure sémiotique

L'étude de la structure sémiotique d'un morceau de musique repose sur trois types d'analyse du discours musical qui conditionnent les hypothèses sur la localisation des frontières structurelles :

- *l'analyse morphologique*, qui porte sur la structure interne des blocs structurels,
- *l'analyse paradigmatique*, qui analyse les relations entre blocs structurels d'une même classe et les oppositions entre blocs sémiotiques de classes différentes,
- *l'analyse syntagmatique*, qui étudie les blocs structurels par leurs voisins au sein du morceau.

En pratique, ces trois analyses se complètent et doivent être effectuées conjointement. Elles amènent en général à identifier à la fois les frontières structurelles et les étiquettes des blocs structurels. Mais, avant de détailler en quoi elles consistent, il est nécessaire de déterminer les strates d'information musicales sur lesquelles travailler.

3.5.1 Strates d'information et propriétés structurantes

Contrairement à la parole et comme nous l'avons défini dans la partie 1.5.2, la musique peut être décrite selon plusieurs strates d'information musicales. Cependant, toutes les strates ne sont pas forcément utiles pour l'étude de la structure d'un morceau donné. Pour un certain nombre de morceaux, les blocs structurels sont construits sur

quelques progressions harmoniques qui se répètent au cours du morceau avec une forte variabilité de la ligne mélodique, alors que d'autres morceaux peuvent être construits sur un seul cycle harmonique du début à la fin, la ligne mélodique étant la seule strate permettant de distinguer les différents blocs.

La pertinence d'une strate du point de vue de l'analyse structurelle peut ainsi s'évaluer en considérant l'évolution de son contenu, c'est-à-dire de ses *propriétés*. Une strate dont les propriétés sont constantes ou cycliques au cours du temps (répétition d'un même accord ou d'une suite d'accords sur l'ensemble du morceau) ou au contraire complètement aléatoires (aucune suite d'accords ne se répète pendant le morceau) ne nous permettra pas d'émettre des hypothèses sur la nature et les relations entre les blocs structurels. La ou les strates utiles à l'analyse de la structure pour un morceau donné sont celles qui se répètent ni trop ni trop peu. Nous les appelons *strates structurantes*.

Comme le montrent les exemples précédents, une strate structurante pour un morceau donné peut se comporter différemment pour un autre et perdre cette qualité.

3.5.2 Indices structurants

Types d'indices structurants L'analyse sémiotique d'un morceau de musique permet d'effectuer un ensemble d'hypothèses sur la localisation des frontières des blocs structurels d'où découle l'inférence d'un ensemble d'*indices structurants*. Il peut s'agir de comportements particuliers de certaines propriétés des strates structurantes ou d'événements pouvant être saillants observés au cours du morceau, comme nous le détaillons dans le paragraphe suivant.

Cependant, tous ces indices peuvent ne pas correspondre de façon univoque à une frontière structurelle, et les indices structurants ayant été utiles pour un morceau de musique particulier peuvent ne plus être pertinents pour un autre.

Dans cette optique, il est nécessaire de faire la distinction entre d'une part les indices structurants *a priori*, qui correspondent à ceux qui peuvent être pressentis comme ayant de fortes chances d'être structurants (avant toute écoute du morceau), et d'autre part les indices *a posteriori* qui marquent effectivement les frontières des blocs, et nécessite que la structure sémiotique soit connue (à l'issue des trois analyses décrites dans les parties suivantes).

Les indices *a priori* sont utiles dans un premier temps afin de proposer des hypothèses relatives aux pulsations structurelles des morceaux. Cependant, elles peuvent, dans certains cas, être revues au cours du processus d'annotation. On peut remarquer notamment que les frontières structurelles des morceaux de musique conventionnels coïncident assez souvent à des ruptures de timbre, c'est-à-dire à l'apparition ou la disparition d'instruments survenant au cours du morceau ainsi qu'à la manière dont ils sont joués. En revanche, de telles ruptures peuvent être également observées à d'autres endroits, par exemple au milieu d'un bloc structurel lorsqu'il présente un enrichissement instrumental progressif.

Recherche d'indices structurants Dans le cadre de la musique conventionnelle, les différentes organisations temporelles tendent à montrer des comportements quasi-cycliques, avec le retour à intervalles réguliers (et non forcément périodiques) d'une strate particulière à un état particulier ou à un ensemble d'états. Par exemple, les motifs rythmiques montrent généralement une récurrence à court-terme qui participe à l'organisation à moyen-terme du morceau de musique, les mélodies ont tendance à

revenir à la tonique ou à faire intervenir des intervalles particuliers selon la position des le bloc structurel, ou encore des séquences d'accords particulières qui se terminent sur des progressions identifiables (cadences harmoniques).

Il existe, au sein des morceaux de musique conventionnels, des instants pour lesquels les strates de mélodie, de l'harmonie et du rythme montrent une convergence de leurs fins de cycles respectifs. Cette convergence constitue un *indice* d'une frontière potentielle entre deux blocs voisins. Les instants de convergence prennent des formes très variables étant donné qu'ils peuvent être signalés à l'intérieur du contenu musical par des combinaisons très diverses d'*indices structurants*, comme un motif rythmique particulier combiné avec le retour à une note ou un accord spécifique, la fin d'un système de rimes dans les paroles ou la conclusion d'une *carrure*[3] avec un effet sonore récurrent.

Itérabilité, suppressibilité Même si ces indices et leurs combinaisons sont en partie spécifiques à chaque genre musical, ils varient généralement d'un morceau à l'autre et sont présents dans tous les genres de musique conventionnelle. Leur identification fait partie de l'analyse empirique réalisée par l'annotateur.

Dans notre approche, la cyclicité joue un rôle central pour l'identification des blocs structurels, par l'intermédiaire des propriétés suivantes :

1. *itérabilité* : un bloc structurel peut être bouclé pour former un segment musical cohérent de plus grande envergure,
2. *suppressibilité* : un bloc structurel peut être enlevé du morceau de musique sans créer la perception d'une discontinuité du flux musical global.

On peut illustrer ces propriétés en considérant un signal périodique. Chacune de ses périodes peut être répétée à l'infini et peut être supprimée du signal sans distordre sérieusement l'organisation du signal modifié. Ceci se généralise conceptuellement à des processus quasi-cycliques.

L'aptitude de l'auditeur à identifier des segments itérables et suppressibles dans un morceau de musique est un point clé de l'analyse proposée et ne nécessite pas que l'annotateur soit capable d'exprimer avec des termes musicologiques les propriétés liées aux indices structurels.

3.5.3 Analyse morphologique : le modèle système - contraste

L'analyse morphologique s'effectue à l'échelle dite *morpho-syntagmatique*, située en dessous de l'échelle des blocs sémiotiques. Elle vise à caractériser chaque bloc structurel par les relations que ses éléments constitutifs partagent entre eux.

Nous faisons l'hypothèse que tout bloc structurel est construit autour d'un ensemble d'en général quatre *éléments morphologiques*, qui forment un *système carré* "porteur" de la structure. Ces éléments, typiquement constitués de deux mesures musicales chacun, sont reliés par une matrice de relations simples, généralement de taille 2×2. Le quatrième élément peut cependant se distinguer de la séquence logique formée par les trois premiers sur certaines strates d'information, ce qui conduit à une forme de contraste[4].

3. "procédé de construction d'une phrase musicale qui divise celle-ci en deux, trois ou n parties de même taille (2, 4 ou 8 mesures)" [AdM01]

4. Notre expérience d'annotation nous a montré que la morphologie des blocs structurels pouvait généralement être expliquée en référence à des radicaux carrés [BDSV12c].

Ce modèle est dénommé *système - contraste* (S&C) [BDSV12b, BDSV12c] et peut se formuler de la manière suivante :

$$af(a)g(a)\gamma(g(f(a)))$$

où a est l'élément morphologique d'*amorce*, f et g sont les relations entre les éléments du système porteur de la structure et γ une fonction de *contraste* qui représente la disparité relative du quatrième élément. Le S&C peut se manifester sur plusieurs strates musicales à la fois. Ceci contribue à la cohérence musicale des blocs structurels et est très utile pour l'identification des frontières structurelles.

Un S&C peut être résumé à l'aide du quadruplet : a, f, g, γ. Cependant, pour un grand nombre de blocs, il s'avère que f ou g (voire les deux) correspondent à l'identité (id), ce qui entraîne l'observation d'un ensemble de motifs morphologiques très courants comme $aaaa, abab, aabb$ (pour $\gamma = id$) ou $aaab, abac, aaba, aabc$ (pour $\gamma \neq id$). Ces motifs peuvent être directement étendus à des fonctions proches de l'identité : $aaa'b, aba'c, aa'bc, ...$ Il a déjà été fait mention de tels motifs comme la forme *period* (abac) et la forme *sentence* (aabc) dans la littérature [Cap98].

Les blocs constitués de plus de quatre éléments morphologiques peuvent être observés, notamment dans les morceaux de blues avec 6 éléments morphologiques. Une description détaillée des différents cas de figure est proposée dans [BDSV12c].

Cette analyse influence à la fois le choix de la granularité choisie pour l'annotation de la structure sémiotique, la localisation des frontières des blocs ainsi que leur étiquetage.

Ainsi la notion d'équivalence entre blocs utilisée pour l'attribution des étiquettes aux blocs structurels s'appuie sur le modèle S&C. Nous associons un ensemble de blocs à une même classe d'équivalence s'ils possèdent :

- les mêmes strates structurantes,
- la même amorce et les mêmes relations sur leurs trois premiers éléments morphologiques.

Comme le quatrième élément morphologique peut ou non contraster vis à vis de la logique instaurée par les trois premiers, la variabilité de son contenu musical ne semble donc pas utile pour comparer deux blocs structurels. Une classe d'équivalence particulière pourra donc se définir à l'aide d'un triplet $\{a, f, g\}$.

3.5.4 Analyse paradigmatique

L'analyse de l'organisation paradigmatique du discours musical correspond à l'analyse des relations entre blocs d'une même classe et des oppositions entre blocs de classes différentes. On peut rapprocher cette analyse de celle de Ruwet [Ruw87] qui consiste en pratique à rechercher des motifs musicaux répétés au cours du morceau. La répétition peut être exacte ou approchée, c'est-à-dire concerner deux portions de morceau qui sont identiques, similaires, ou plus généralement faciles à relier l'un à l'autre à partir de transformations simples comme la transposition, la modification de l'instrumentation, l'ajout d'un motif mélodique superposé au bloc, la troncature ou l'insertion d'un affixe.

Comme on le verra aussi dans le cadre de l'analyse syntagmatique, la localisation de tels paradigmes ne coïncide pas toujours de manière univoque avec les frontières structurelles : ils constituent des indices supplémentaires pour l'annotation de ces frontières.

L'analyse paradigmatique est utile à l'étiquetage sémiotique des blocs structurels : on associe une même étiquette sémiotique aux blocs semblables du point de vue des strates structurantes du morceau. Celle-ci est complétée par l'analyse morphologique

du paragraphe précédent.

3.5.5 Analyse syntagmatique

L'étiquetage sémiotique des blocs structurels, qui repose principalement sur les analyses morphologique et paradigmatique, nécessite parfois de considérer son patron structurel par lequel on étudie la position et le contexte de chaque bloc dans le morceau. On considère que deux blocs structurels ont des chances d'appartenir à la même classe d'équivalence et donc d'être associés à une même étiquette s'ils apparaissent dans des contextes semblables, c'est-à-dire s'ils sont situés à droite et/ou à gauche de segments similaires du point de vue sémiotique. Par exemple, dans la séquence ABxDAByDECDCDD, il est plus vraisemblable de regrouper x et y au sein d'une même classe que dans la séquence ABxDyBCDECDCDD. Bien qu'intéressante, l'introduction de considérations syntagmatiques pour l'étiquetage des blocs structurels ne fait pas l'objet de cette thèse. Le lecteur pourra se référer à [BDSV12a] pour une discussion plus fournie sur ce sujet.

3.6 Méthodologie pratique d'annotation

Les concepts et principes présentés précédemment permettent de définir une méthodologie d'annotation de la structure sémiotique d'un morceau de musique, qui peut être résumée par les trois étapes suivantes :

1. Une première écoute globale du morceau permet
   - l'identification des dimensions structurantes,
   - la recherche d'indices structurants,
   - l'hypothèse d'une ou plusieurs pulsations structurelles.

   L'organisation du début et de la fin du morceau sont souvent assez atypiques en comparaison du reste du morceau. Ils peuvent, dans un premier temps, être écartés de l'analyse du morceau.
2. Une écoute approfondie des dimensions structurantes permet ensuite à l'annotateur de segmenter le morceau, et d'étiqueter les segments obtenus en fonction
   - de l'organisation interne de chaque bloc (analyse morphologique),
   - des correspondances à long terme (analyse paradigmatique),
   - des étiquettes de leurs voisins en cas d'ambiguïté (analyse syntagmatique).
3. Si la décomposition obtenue apparaît trop complexe pour rendre compte de l'organisation du morceau, une nouvelle hypothèse de pulsation structurelle peut être émise afin de reprendre l'étape précédente.

Une explication plus détaillée de cette méthodologie ainsi qu'un cas d'étude sont présentés dans [BDSV11].

3.7 Annotations structurelles produites

La méthodologie proposée a permis l'annotation de 383 morceaux de musique [BDSV11], issus de plusieurs bases différentes. L'annotation de 100 morceaux supplémentaires est en cours.

Base RWC Pop Il s'agit des annotations de la RWC Popular Music Database qui consiste en 100 morceaux de pop produits pour la recherche et "dans le style des chansons des *hit charts* américains des années 80 et de celui de la musique populaire des *hit charts* japonais des années 90" [GHNO02] [5]. Une version de ces annotations ne comprenant que les frontières structurelles a notamment été utilisée dans le cadre des campagnes d'évaluation MIREX 2010, 2011 et 2012.

Base Quaero Cette base est composée de 159 titres sélectionnés par l'IRCAM, dans le cadre du projet Quaero pour l'évaluation d'algorithmes d'estimation de la structure :

Quaero 2009 : 69 titres

Quaero 2010 : 45 titres

Quaero 2011 : 45 titres

Les morceaux de la base durent en moyenne 4 minutes. Elle comporte des morceaux de genres variés (pop, rock, hard rock, rap, électro), dont les artistes sont en majorité d'origine américaine ou anglaise. La liste des morceaux de la base est présentée en annexe A [6].

Base Eurovision Cet ensemble regroupe 124 titres correspondant à la version studio des chansons ayant participé aux demi-finales ou aux finales du concours Eurovision de la chanson des années 2008, 2009 et 2010. Celles-ci figurent sur les compilations officielles dont les références sont données ci-dessous.

2008 Belgrade, ref # 5 099921 699726 : 43 titres

2009 Moscou, ref # 5 099969 968020 : 42 titres

2010 Oslo, ref # 5 099964 171722 : 39 titres

Les morceaux d'Eurovision ont leur durée limitée à un maximum de 3 minutes par les règles du concours. Ils ont certaines particularités sous l'influence du format du concours, ainsi que de son public. Cependant, ces titres couvrent une certaine variété de langues et une diversité de sous-genres de pop européenne.

Base RWC Genre Cet ensemble est constitué des 100 morceaux de la base RWC Music Genre Database [GHNO02]. Elle est divisée en 10 catégories de genre, et 33 sous-catégories de genres. Cet ensemble est en cours d'annotation et n'est pas encore accessible.

Les annotations des trois premières bases sont disponibles à l'adresse

`http://musicdata.gforge.inria.fr`

ainsi que sur la plateforme d'écoute et de visualisation en ligne

`http://metissannotation.irisa.fr`

Cette plateforme a été créée dans le but de discuter et de diffuser la méthodologie et les annotations obtenues.

5. La liste des morceaux est disponible à l'adresse `http://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-p.html`

6. Contacter le département Analyse-Synthèse de l'IRCAM pour plus de renseignement concernant cette base.

3.8 Résumé du chapitre

Ce chapitre résume les contributions apportées pendant cette thèse sur la spécification de la structure des morceaux de musique et décrit les concepts, axiomes et principes méthodologiques qui en sont issus.

Le problème est formulé du point de vue de la linguistique structuraliste en considérant que la structure est un système dont les éléments, les segments structurels (ou *blocs*), se définissent essentiellement par l'ensemble des relations qui les lient. Cette structure est supposée elle-même être issue d'un modèle de structure, le *patron structurel*, qui est constitué d'une séquence de blocs prototypiques à partir desquels les blocs structurels sont réalisés. Les blocs prototypiques sont en nombre limité et de taille comparable. Les blocs structurels peuvent être décomposés en un *radical* éventuellement soumis à un ensemble de distorsions temporelles, les *affixes* et les troncatures. Il découle du patron structurel et de la comparabilité des blocs une hypothèse de *pulsation structurelle*, c'est-à-dire l'existence d'un nombre limité de tailles de radicaux et d'une faible fluctuation autour de ces valeurs typiques.

Nous présentons ensuite plusieurs principes pour l'analyse de la structure par des annotateurs humains. L'écoute initiale du morceau permet d'effectuer un ensemble d'hypothèses sur les frontières structurelles et sur la pulsation structurelle. Ceci permet de déterminer quelles sont les strates musicales pertinentes pour expliquer la structure du morceau. La désambiguisation des hypothèses de segmentation structurelle passe alors par la réalisation conjointe de l'analyse morphologique (relative à l'organisation interne des blocs et s'appuyant sur le modèle *système-contraste* proposé), paradigmatique (étude des répétitions à l'échelle du morceau), et syntagmatique (étude de l'agencement des blocs structurels) du morceau, ce qui peut conduire à réévaluer ces hypothèses. Cette méthodologie d'annotation a donné lieu à la production de plusieurs bases d'annotations totalisant un demi-millier de morceaux, destinées à être mises à la disposition de la communauté scientifique.

Chapitre 4

Approches pour l'estimation de la structure sémiotique

Nous introduisons plusieurs approches afin d'estimer automatiquement la structure sémiotique. Celles-ci découlent de l'étude de l'état de l'art du chapitre 2 et des développements méthodologiques du chapitre 3. Nous nous focalisons tout d'abord sur l'estimation des frontières structurelles dans la partie 4.1, puis sur l'estimation des étiquettes sémiotiques en supposant les frontières connues dans la partie 4.2.

4.1 Approches pour l'estimation des frontières structurelles

Les considérations méthodologiques du chapitre 3 ont permis de préciser la structure à long terme que nous allons chercher à retrouver automatiquement. L'étude de cette structure est fondée sur l'analyse conjointe d'un ensemble de strates musicales dites structurantes, et repose sur l'hypothèse de pulsation structurelle. C'est dans cette optique que nous développons ici une approche pour estimer les frontières structurelles des morceaux de musique, en utilisant conjointement plusieurs critères audio sous une contrainte dite de *régularité structurelle*. Cette contrainte a pour objectif de favoriser les blocs structurels de taille proche de la pulsation structurelle.

Nous formulons tout d'abord le problème de la segmentation comme un problème d'optimisation d'un *coût de segmentation*. Celui-ci permet de faire clairement apparaître les contributions liées aux critères et contraintes considérés. Nous développons ensuite notre approche *multicritère* en introduisant un cadre probabiliste pour l'expression et la combinaison des critères de segmentation avant de présenter les différents critères que nous considérons dans notre étude. Un critère morphologique basé sur le modèle système - contraste est introduit à titre exploratoire. Enfin, nous formulons une contrainte de régularité incorporant l'hypothèse de pulsation structurelle à la recherche des frontières structurelles.

4.1.1 Cadre général pour la segmentation

Notre panorama de l'état de l'art du chapitre 2 a permis de mettre en valeur le fait que l'analyse du contenu musical d'un morceau de musique est effectuée à l'aide de critères audio et de contraintes structurelles, dans l'optique d'aboutir à une segmenta-

tion unique. Nous proposons ici de formuler le problème de la segmentation structurelle à l'aide un processus d'optimisation d'un coût global qui prend en compte ces deux composantes.

Soit X un morceau de musique, qui peut être décrit par une séquence de T vecteurs de descripteurs. On définit une segmentation $S = \{s_k\}_{1 \leq k \leq K}$ de X comme une séquence de K intervalles $s_k = [t_k, t_{k+1}[$.

On considère que chaque segmentation possède un certain coût, lié à la fois au respect des critères de caractérisation des blocs et ainsi qu'aux contraintes structurelles choisies. La segmentation structurelle recherchée est celle de coût minimal. On définit la fonction de *coût de segmentation* suivante :

$$C(S) = \sum_{k=1}^{K} \Gamma(s_k) \tag{4.1}$$

avec

$$\Gamma(s_k) = (1 - \lambda)\Phi(s_k) + \lambda\Psi(s_k). \tag{4.2}$$

$\Phi(s_k)$ est une fonction de coût liée aux critères audio. Elle prend des valeurs faibles lorsque la séquence de descripteurs associée à s_k est susceptible de correspondre à un bloc structurel, suivant la ou les caractérisations choisies.

$\Psi(s_k)$ est une fonction de coût liée aux contraintes structurelles. Elle permet de prendre en compte les hypothèses émises sur la nature de la structure, par exemple sur la taille et le nombre de blocs structurels désirés. $\Psi(s_k)$ prendra des valeurs faibles si le bloc s_k présente des propriétés en accord avec ces hypothèses.

λ est un paramètre de compromis entre ces deux coûts, qui varie entre 0 et 1.

Cette formulation permet de faire apparaître clairement les différentes composantes du problème de la segmentation, et peut prendre en compte une grande variété de critères et de contraintes, pourvu qu'ils soient exprimés conformément aux spécifications ci-dessus. Elle est donc générique, dans le sens où il est possible d'exprimer les différentes approches de segmentation structurelle dans ce cadre.

4.1.2 Critères de segmentation structurelle

L'état de l'art a mis en valeur l'utilisation de plusieurs critères audio individuels pour l'estimation des frontières structurelles. Il s'agit en général de la détection d'indices structurants *a priori* (tels que définis dans la partie 3.5.2) et la diversité des manifestations de la structure sémiotique entraîne de ce fait une efficacité variable de ces critères selon les morceaux de musique.

Par ailleurs, la structure musicale s'exprime en général sur plusieurs niveaux simultanément (niveaux de surface et niveaux structurels), ce qui nous amène à considérer, dans la partie 4.1.2.1 des combinaisons de critères plutôt que l'utilisation d'un critère unique. Cette approche est en outre motivée par les considérations méthodologiques du chapitre 3, qui considèrent conjointement plusieurs analyses (morphologique, paradigmatique et syntagmatique) afin de déterminer la structure sémiotique des morceaux.

Les méthodes d'estimation de structure présentées dans cette thèse se répartissent dans ces différentes catégories. Nous formulons d'abord trois critères dans un cadre probabiliste : un critère de rupture d'homogénéité, un critère de rupture de répétition et un critère de détection d'événements.

Le premier et le troisième opèrent sur des niveaux de surface, et entrent dans la catégorie des méthodes de détection d'indices structurants *a priori*. Le critère de rupture d'homogénéité consiste à détecter des ruptures de timbre qui tendent à coïncider avec les frontières structurelles des morceaux de musique conventionnelle. Il en va de même pour le critère de détection d'événements qui recherche les segments courts et atypiques, susceptibles de caractériser les percussions ou les événements exceptionnels fréquemment situés de fin de segment.

Le critère de rupture de répétition se rattache à une analyse paradigmatique conformément à la méthodologie d'annotation de la structure sémiotique ; en ce sens, il entre dans la catégorie des critères opérant au niveau structurel.

La volonté d'utiliser plusieurs critères pour l'estimation des frontières structurelles pose la question de leur combinaison. Nous considérons dans cette thèse leur combinaison par l'intermédiaire d'une somme pondérée, en les ayant préalablement exprimés dans un même cadre.

Nous introduisons également, dans la partie 4.1.2.4 un critère morphologique expérimental, lui aussi relevant du niveau structurel. Il est cependant exploité séparément par rapport aux critères précédents car il n'est pas formulé dans un cadre probabiliste.

4.1.2.1 Cadre probabiliste pour la combinaison de critères audio

Nous avons besoin d'un cadre nous permettant de formuler toute une variété de critères de segmentation en vue de combiner des quantités de même nature. C'est ce qui nous a amené à nous intéresser au logarithme du rapport de vraisemblance généralisé log(RVG). Il s'agit d'une statistique de test issue de la théorie de la décision. Elle se définit comme suit.

Soit $Y = \{Y_t\}_{1 \leq t \leq n}$ un processus aléatoire à valeurs dans $\mathbb{R}^d$, $d \in \mathbb{N}$. Les variables aléatoires Y_t qui le composent sont supposées indépendantes et telles que leur densité de probabilité appartient à la famille $p(.|\theta), \theta \in \Theta$, Θ désignant l'espace des paramètres considérés. Notons $y = \{y_t\}_{1 \leq t \leq n}$ une séquence d'observations (ou *réalisation*) issue de Y, et Θ_0, Θ_1 deux sous ensembles de Θ disjoints et non réduits à un singleton. Le RVG est une quantité utilisée pour effectuer une décision entre deux hypothèses *composites* H_0 et H_1 sur la loi de paramètre θ^* à l'origine de la production de la réalisation y, c'est-à-dire :

- $H_0 : \theta^* \in \Theta_0$
- $H_1 : \theta^* \in \Theta_1$

Le RVG est défini comme suit :

$$\text{RVG} = \frac{P(y|H_1)}{P(y|H_0)} = \frac{\sup_{\theta \in \Theta_1} p(y|\theta)}{\sup_{\theta \in \Theta_0} p(y|\theta)} \tag{4.3}$$

Il s'agit d'un rapport entre deux maxima de vraisemblance : la probabilité que la réalisation y soit produite par la loi de paramètre θ_1^* qui la modélise le mieux parmi les lois de paramètre $\theta \in \Theta_1$, et la probabilité que y soit produite par la loi de paramètre θ_0^* qui la modélise le mieux parmi les lois de paramètre $\theta \in \Theta_0$. Le RVG se distingue d'un rapport de vraisemblance classique dans lequel Θ_0 et Θ_1 sont des singletons : H_0 et H_1 sont alors des hypothèses *simples*, et les lois de production de y sont connues *a priori*.

En pratique, y correspond à une séquence de vecteurs de descripteurs, et les hypothèses H_0 et H_1 sont antagonistes. On obtient ainsi un log(RVG) maximal lorsque l'hypothèse H_1 est vraisemblable, qui devient minimal lorsque H_0 l'est.

Ce cadre est notamment utile afin d'exprimer toute une variété de critères de segmentation.

4.1.2.2 Critères audio formulés à l'aide d'un RVG

Nous exprimons trois critères de détection des frontières structurelles par l'intermédiaire d'un RVG : un critère de rupture d'homogénéité, un critère de détection d'événements localisés et un critère de rupture de répétition.

Présentation des trois critères Le critère de rupture d'homogénéité consiste à comparer, pour chaque instant considéré, un modèle gaussien issu des descripteurs précédant cet instant par rapport à celui issu des descripteurs qui le suivent, pour un voisinage particulier. L'hypothèse H_0 correspond à l'absence de frontière structurelle à l'instant considéré et se traduit par une forte ressemblance entre les deux modèles. L'hypothèse antagoniste H_1 renvoie à la présence d'une frontière et est associée à une faible ressemblance entre les modèles (c'est-à-dire une rupture d'homogénéité).

Le critère de rupture de répétition évalue pour chaque instant si la séquence de descripteurs contenue dans un voisinage fixé autour de cet instant se retrouve ailleurs dans le morceau. La recherche de répétition passe par la comparaison des séquences de descripteurs modélisées par des lois gaussiennes à variance fixe. L'hypothèse H_0 correspond au fait que cette séquence est répétée intégralement. L'hypothèse H_1 correspond au fait que les parties passée et future de la séquence par rapport à l'instant considéré sont au mieux répétées mais disjointes au cours du morceau. La fenêtre d'analyse utilisée, constituée des parties "passé" et "futur", est ainsi du même type que celle du critère de rupture d'homogénéité. Cependant, les hypothèses considérées sont différentes.

Le critère de détection d'événement est basé sur l'observation de la présence d'*événements localisés* à la fin des blocs structurels. On considère qu'un événement est une perturbation importante du flux audio sur une durée de l'ordre de la seconde préparant l'auditeur à l'arrivée d'un nouveau bloc, et correspond à un indice structurant de type a priori (*cf.* partie 3.5.5). Il s'agit par exemple d'un roulement de batterie, de l'ajout d'effets sonores ou au contraire d'un bref appauvrissement de l'instrumentation du morceau pouvant aller jusqu'au silence. Le critère de détection d'événement consiste en la comparaison des deux modèles gaussiens respectivement issus des descripteurs associés à l'événement (contenus dans un petit voisinage autour de l'instant considéré) et de ceux de son environnement (contenus dans un voisinage plus large ne comprenant pas celui de l'événement). Il reprend les mêmes hypothèses que le critère de rupture d'homogénéité : H_0 correspond à l'absence de frontière structurelle ce qui est traduit par une forte ressemblance entre les deux modèles, et l'hypothèse de présence d'une frontière H_1 correspond au cas contraire. En revanche, soulignons qu'il diffère du critère d'homogénéité d'un point de vue conceptuel : il s'agit d'identifier la présence d'un événement au regard de son environnement passé et futur, et non pas de localiser une rupture d'homogénéité en comparant le passé au futur.

Formulation du critère de rupture d'homogénéité Ce critère a déjà été formulé par un RVG dans le cadre de la segmentation parole-musique [SBB00] et se rapproche conceptuellement du critère de nouveauté de Foote décrit dans la partie 2.6.

Soit $y^0 = \{y_t^0\}_{1 \leq t \leq 2N}$ une séquence d'observations à valeurs dans $\mathbb{R}^d$, avec d et $N \in \mathbb{N}$. Notons $y^1 = \{y_t^1\}_{1 \leq t \leq N}$ la première moitié de la séquence y^0, et $y^2 = \{y_t^2\}_{1 \leq t \leq N}$

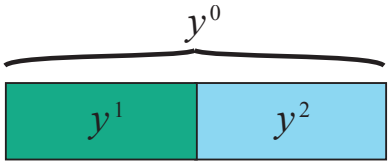

FIGURE 4.1 – Fenêtre d'analyse associé aux critère de rupture d'homogénéité et de répétition, centrée sur l'instant courant. Celle-ci est constituée de deux demi-fenêtres successives de taille égale. La première est associée aux N descripteurs précédant l'instant courant, la seconde est associée aux N descripteurs qui le suivent.

sa seconde moitié, comme illustré dans la figure 4.1.

Les hypothèses sont définies comme suit :

- H_0 : la séquence y^0 est issue d'une seule distribution de probabilité (hypothèse de non-rupture d'homogénéité en $t = N$),
- H_1 : les séquences y^1 et y^2 sont respectivement issues de deux distributions de probabilité distinctes (hypothèse de rupture d'homogénéité en $t = N$).

En pratique, les paramètres des distributions utilisées lors du calcul du log(RVG) correspondent aux estimateurs du maximum de vraisemblance. On note G_0, G_1 et G_2 les distributions respectivement associées aux séquences y^0, y^1 et y^2. Le critère s'écrit :

$$\phi_H = \log(\text{RVG}) = \log \frac{P(y^1|G_1)P(y^2|G_2)}{P(y^0|G_0)} = \log P(y^1|G_1) + \log P(y^2|G_2) - \log P(y^0|G_0) \tag{4.4}$$

Dans le cas de descripteurs numériques modélisés par des distributions gaussiennes ce critère peut s'exprimer comme

$$\phi_H = N(\log(\det(\Gamma_0)) - \frac{\log(\det(\Gamma_1)) + \log(\det(\Gamma_2))}{2}) \tag{4.5}$$

où Γ_0, Γ_1 et Γ_2 sont les matrices de covariance associées aux descripteurs des séquences y^0, et y^1 et y^2. Cette quantité est maximale lorsque y^1 et y^2 suivent des distributions distinctes. Le détail du calcul permettant d'obtenir ce résultat est donné dans l'annexe B.

Formulation du critère de rupture de répétition Soit $y = \{y_t\}_{1 \leq t \leq T}$ une séquence d'observations issue du processus aléatoire Y à valeurs dans $\mathbb{R}^d$. Soit $y^0 = \{y_t^0\}_{1 \leq t \leq 2N}$ une portion de cette séquence, avec $0 < 2N << T$. On note $y^1 = \{y_t^1\}_{1 \leq t \leq N}$ la première moitié de y^0, et $y^2 = \{y_t^2\}_{1 \leq t \leq N}$ sa seconde moitié comme dans la figure 4.1.

Nous utilisons les hypothèses suivantes afin de formuler un critère de rupture de répétition à l'aide du log(RVG) :

- H_0 : la séquence y^0 se répète ailleurs au sein de la séquence y, c'est-à-dire qu'il existe une suite d'éléments de y proche de y^0 (hypothèse de non-rupture de répétition en N).
- H_1 : les séquences y^1 et y^2 se répètent ailleurs dans la séquence y mais n'apparaissent pas de manière successive ailleurs que dans y^0 (hypothèse de rupture de répétition en N).

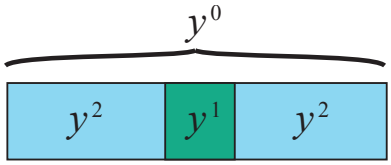

FIGURE 4.2 – Fenêtre d'analyse associé au critère de détection d'événement localisé centrée sur l'instant courant. Celle-ci est constituée d'une petite fenêtre de taille $2L$ concernant le proche voisinage de l'instant courant par la séquence de descripteurs y^1, et d'une fenêtre composite de taille totale $2(N-L)$ concernant son environnement passé et futur par y^2.

Les maxima de vraisemblance du RVG impliquent ici que chaque séquence y^0, y^1, y^2 est modélisée par la séquence qui lui est la plus semblable dans y. On les note $g^0 = \{g_t^0\}_{1\leq t\leq 2N}$ pour y^0, $g^1 = \{g_t^1\}_{1\leq t\leq N}$ pour y^1, et $g^2 = \{g_t^2\}_{1\leq t\leq N}$ pour y^2.

Le log(RVG) s'écrit :

$$\phi_R = \log(\text{RVG}) = \log\left(\frac{P(y^1|g^1)P(y^2|g^2)}{P(y^0|g^0)}\right) = \log\left(P(y^1|g^1)\right) + \log\left(P(y^2|g^2)\right) - \log\left(P(y^0|g^0)\right) \quad (4.6)$$

Dans le cas des descripteurs numériques, on suppose que Θ_0 correspond à l'espace des paramètres des lois de probabilité gaussienne à variance fixe, et que Θ_1 correspond à l'espace des paramètres des paires de lois de probabilité gaussienne à variance fixe, produisant toute sous-séquence de y de taille N. Ceci permet d'aboutir à la formule analytique suivante :

$$\phi_R \propto -\sum_{t=1}^{N} \|g_t^1 - y_t^1\|^2 - \sum_{t=1}^{N} \|g_t^2 - y_t^2\|^2 + \sum_{t=1}^{2N} \|g_t^0 - y_t^0\|^2 + constante \quad (4.7)$$

Le lecteur pourra se référer à l'annexe B pour avoir le détail du calcul permettant d'aboutir à cette forme.

Formulation du critère de détection d'événement localisé Soit $y^0 = \{y_t\}_{1\leq t\leq 2N}$ une séquence d'observations, avec $N \in \mathbb{N}$. Soit $L < N$, on note $y^1 = \{y_t\}_{N-L+1\leq t\leq N+L}$ la sous-séquence d'observations associée à l'événement de taille $2L$ trames, et $y^2 = \{y_t\}_{t\in[\![1,N-L]\!]\cup[\![N+L+1,2N]\!]}$ la sous-séquence d'observations associée à son environnement, passé et futur, de taille totale $2N - 2L$ trames comme illustré dans la figure 4.2.

Ce critère de détection d'événement reprend les hypothèses du critère de rupture d'homogénéité décrit ci-dessus :

- H_0 : la séquence y^0 est issue d'une seule distribution de probabilité (hypothèse d'absence d'événement localisé en N)
- H_1 : les séquences y^1 et y^2 sont respectivement issues de deux distributions de probabilité distinctes (hypothèse de présence d'événement localisé en N)

Dans le cas des descripteurs numériques, si l'on note Γ_0, Γ_1 et Γ_2 les matrices de covariance respectives de y^0, y^1 et y^2, le critère de détection d'événement localisé s'écrit,

en procédant de la même manière que pour le critère d'homogénéité :

$$\phi_E = \log(\text{RVG}) = N \log\left(\det\left(\Gamma_0\right)\right) - L \log\left(\det\left(\Gamma_1\right)\right) - (N - L) \log\left(\det\left(\Gamma_2\right)\right) \quad (4.8)$$

4.1.2.3 Combinaison des critères audio

La combinaison de plusieurs critères audio permet de considérer simultanément plusieurs caractéristiques de la structure recherchée afin d'en estimer les frontières.

Cette problématique entre dans le cadre de l'optimisation multicritère, qui constitue l'une des branches de la théorie de la décision. S'il est théoriquement simple de rechercher une solution optimale sur la base d'un seul critère, le problème se complexifie lorsque plusieurs critères entrent en jeu : non seulement plusieurs solutions alternatives peuvent optimiser le problème, mais la notion même d'*optimalité* peut prendre plusieurs définitions ([Ehr05], p7). La plus répandue est celle d'*efficacité* ou optimalité au sens de Pareto : une solution est dite *efficace* s'il n'existe pas d'autre solution capable d'améliorer (au sens large) l'ensemble des critères considérés (pour une définition formelle, voir [Ehr05], p24). Un grand nombre de méthodes ont été développées pour la recherche des solutions efficaces [MA04].

Dans cette thèse, nous utilisons une somme pondérée afin de combiner plusieurs critères partageant la même formulation. Ce procédé appartient aux méthodes dites *weighted global criterion* visant à se ramener à l'optimisation d'un critère unique, obtenu par combinaison des critères considérés. L'étude d'autres méthodes de combinaison constitue une piste de recherche que nous ne traitons pas ici.

Assez naturellement, la combinaison linéaire est formulée de la manière suivante :

Soit $\{\phi_i\}_{i=1\ldots I}$ un ensemble de critères audio, $\{\lambda_i\}_{i=1\ldots I}$ un ensemble de poids tel que $\sum_{i=1}^{I} \lambda_i = 1$ et $\lambda_i \geq 0$, le critère audio combiné ϕ_{CL} est obtenu par :

$$\phi_{CL} = \sum_{i=1}^{I} \lambda_i \phi_i. \quad (4.9)$$

Les paramètres de pondération permettent ainsi de prendre en compte leur importance relative en attribuant à un critère un poids d'autant plus élevé qu'il est utile à la localisation des frontières structurelles.

4.1.2.4 Critère morphologique utilisant le modèle système - contraste

Nous proposons ici un critère permettant de considérer l'analyse morphologique introduite dans la partie 3.5.3 pour estimer les frontières structurelles. Ce critère est ici directement formulé sous la forme d'un coût $\Phi(s_k)$ tel qu'il a été conçu et inclus dans le système d'estimation de structure soumis à titre exploratoire à la campagne d'évaluation MIREX en 2012, étudié au chapitre suivant. Cette formulation ne correspond actuellement pas à une formulation de type RVG et ne sera ainsi pas pris en compte dans l'approche multicritère du chapitre 6.

On considère ici que les systèmes sont composés de quatre éléments morphologiques caractérisables par le quadruplet (a, f, g, δ). On fait de plus l'hypothèse que f et/ou g est égale à la fonction *identité* (id) ou à une fonction proche de l'identité (id'). Dans le cas de id, on observe les motifs morphologiques suivants : *aaaa* et *aaab* (f et $g = id$), *aaba* et *aabc* ($f = id$), *abac* ($g = id$). Dans le cas de id', l'amorce a est au moins similaire au deuxième ou au troisième élément morphologique. La fonction δ peut prendre des

formes très diverses, ainsi le quatrième élément morphologique peut être semblable à a ou contraster avec le reste du système auquel il appartient. Nous ne tenons donc pas compte ici du quatrième élément d'un système pour le caractériser. Cependant il nous semble intéressant de considérer le dernier élément morphologique du système précédant afin d'identifier le début du système courant : on observe généralement que cet élément diffère significativement de l'amorce qui lui succède (le cas du préfixe fait exception). Ainsi nous utiliserons une fenêtre d'analyse de taille quatre éléments morphologiques dans l'optique d'identifier le début d'un système commençant sur le deuxième quart de cette fenêtre, comme illustré dans la figure 4.3. Le critère morphologique est formalisé par le coût suivant.

Soit $y^0 = \{y_t^0\}_{1 \leq t \leq 4N}$ une séquence d'observations à valeurs dans $\mathbb{R}^d$, avec d et $N \in \mathbb{N}$. $y^1 = \{y_t^0\}_{1 \leq t \leq N}$ correspond au premier élément morphologique a_0 de la séquence y^0, et $y^2 = \{y_t^0\}_{N+1 \leq t \leq 4N}$ correspond à la séquence des trois éléments suivants : a_1, a_2 et a_3.

Le coût considéré permet de quantifier l'invraisemblance de l'hypothèse selon laquelle une frontière de système coïncide avec l'indice N dans la fenêtre d'analyse, c'est-à-dire que a_0 correspond à la fin d'un système et (a_1, a_2, a_3) correspond à une partie du suivant avec $a_1 = a$. On l'exprime à l'aide de la combinaison linéaire suivante :

$$\Phi_{SC} = \lambda_1 \sigma_{\text{Système}} + \lambda_2 \sigma_{\text{Contraste}} \tag{4.10}$$

avec λ_1 et $\lambda_2 \in \mathbb{R}^+$.

$\sigma_{\text{Système}}$ quantifie la vraisemblance qu'un système débute à l'indice N en évaluant la contribution de a_1 à l'explication de a_2 et a_3. Posons $X_i = \{y^0_{1+iN}, ..., y^0_{N+iN}\}$ pour $i = \{1, 2, 3\}$, et $Z_j = (X_{j+1} - X_1)^2$ pour $j = \{1, 2\}$ on calcule

$$\sigma_{\text{Système}} = \frac{\sum_{l=1}^{N} \min(Z_1(l), Z_2(l))}{||X_1||^2} \tag{4.11}$$

où $||.||$ est la norme euclidienne. Cette quantité permet d'évaluer si X_1 permet de décrire correctement X_2 ou X_3 suivant les différentes dimensions des observations et en tenant compte de l'ordre temporel.

$\sigma_{\text{Contraste}}$ quantifie la vraisemblance que la séquence d'observations précédant l'indice N corresponde à la fin d'un système, et repose sur la différence entre a_0 et a_1. Ceci permet de considérer les cas où soit deux systèmes différents se succèdent, soit deux systèmes identiques se succèdent mais le premier se termine par un contraste. Posons $X_0 = \{y_1^0, ..., y_N^0\}$, on a

$$\sigma_{\text{Contraste}} = \text{cotan}(X_0, X_1) \tag{4.12}$$

qui prend des valeurs faibles lorsque X_0 et X_1 contrastent ($\delta \neq id$) et prend des valeurs élevées dans le cas contraire.

4.1.3 Contrainte de régularité structurelle

Nous proposons d'incorporer au processus de segmentation structurelle l'hypothèse d'existence d'une pulsation structurelle à l'aide d'une contrainte dite *de régularité*.

Le choix de ne considérer qu'une pulsation structurelle est motivée par le fait qu'un grand nombre de chansons est concerné par cette hypothèse. L'observation des annotations des 100 morceaux de la base RWC Pop, considérée dans le chapitre 6, permet

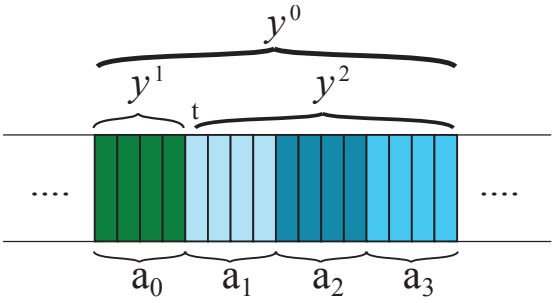

FIGURE 4.3 – Fenêtre d'analyse associée au critère de détection des débuts de systèmes pour un instant t. Elle est composée de quatre sous-fenêtres de taille N associée à quatre éléments morphologiques. Sur l'illustration, $N = 4$.

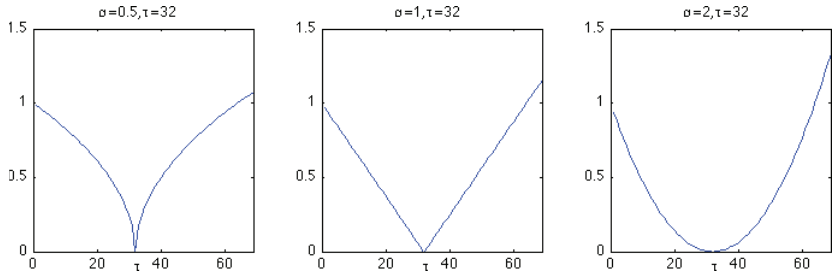

FIGURE 4.4 – Exemples de contraintes de régularité Ψ_α pour $\alpha = \{0.5, 1, 2\}$ et $\tau = 32$ temps (typiquement équivalent à 16 snaps).

de remarquer que 73 morceaux ont une seule taille de radical qui a plus de deux occurrences [1].

L'objectif de la contrainte de régularité est de favoriser les blocs structurels dont la taille est proche de la pulsation structurelle, notée τ. Notre contrainte de régularité est ainsi fondée sur la mesure de la déviation entre la taille des blocs structurels d'une segmentation et τ.

Soit s un bloc de taille m, le coût Ψ associé à s possède les propriétés suivantes :

1. $\Psi(m) = 0$ si $m = \tau$,
2. $\Psi(m) > 0$, en prenant des valeurs d'autant plus élevées que m s'éloigne de τ.

Un grand nombre de fonctions satisfait ces propriétés.

Nous considérons dans cette thèse la famille des fonctions symétriques issues de la valeur absolue à la puissance α :

$$\Psi_\alpha(m) = \left|\frac{m}{\tau} - 1\right|^\alpha \tag{4.13}$$

Le paramètre α contrôle la convexité de la fonction : celle-ci est non-convexe si $0 < \alpha < 1$, et convexe si $\alpha \geq 1$. La figure 4.4 représente Ψ_α pour $\alpha \in \{0.5, 1, 2\}$.

L'utilisation de fonctions non-convexes permet de tolérer les segmentations comportant des blocs très irréguliers, mais peu de blocs légèrement irréguliers. A l'inverse, les fonctions convexes tendent à tolérer les segmentations comprenant davantage de blocs légèrement irréguliers, mais peu de blocs très irréguliers.

1. Tous les radicaux ont une taille supérieure à 3 snaps.

L'étude de l'effet de cette famille de fonction sur la segmentation structurelle fait l'objet de la partie 6.2.

4.1.4 Limites liées à la contrainte de régularité

La contrainte de régularité formulée ci-dessus suppose l'existence d'une seule pulsation structurelle. Cette contrainte est très simple, et l'on peut se demander si une contrainte de régularité plus complexe comportant par exemple plusieurs pulsations structurelles pourrait améliorer la segmentation. Pour ce faire nous étudions l'apport d'une contrainte de régularité "idéale" issue de la distribution des tailles de blocs du morceau courant.

Soient S_r la segmentation de référence d'un morceau de musique, et H_{S_r} l'histogramme normalisé des I tailles de ses blocs $\{\nu_i\}_{1\leq i\leq I}$. La fonction de coût correspondante est définie comme suit :

$$\Psi_r(\nu) = \begin{cases} -\log(\mathrm{H_{S_r}}(\nu)) & \text{si} \quad \nu = \nu_i, 1 \leq i \leq I, \\ +\infty & \text{sinon.} \end{cases} \tag{4.14}$$

Ainsi, les tailles de blocs les plus fréquentes dans l'histogramme sont associées à un faible coût, et à l'inverse, les tailles les moins fréquentes impliquent un fort coût.

Une telle fonction de coût permet ainsi de favoriser les tailles de blocs les plus fréquemment rencontrées au sein du morceau courant. L'évaluation d'un système de segmentation avec ce coût idéal est présentée dans la partie 6.2.2.

4.1.5 Détails d'implémentation

Nous avons formulé un cadre d'étude, un ensemble de critères audio et une famille de contraintes structurelles afin de traiter le problème de la segmentation structurelle. Cette partie porte sur leur mise en œuvre expérimentale.

4.1.5.1 Précision sur le calcul des critères audio

Descripteurs considérés Nous utilisons les descripteurs acoustiques MFCC et les vecteurs de chroma pour les expérimentations associées à l'estimation des frontières structurelles menées dans le chapitre 6.

Dans le cadre de l'extraction des MFCCs le signal est subdivisé en trames de 1024 échantillons. Celles-ci se recouvrent sur 512 échantillons. On extrait de chaque trame les 20 premiers coefficients MFCC. Ce nombre de coefficients est souvent considéré dans le domaine du MIR selon [CVG+08]. Le coefficient d'ordre 0 est considéré : il contient des informations relatives à l'intensité sonore moyenne de la trame [Log00], ce qui nous semble pertinent pour leur utilisation en tant que descripteurs de timbre (cette notion combine des informations liées au spectre et à l'intensité sonore). Les MFCCs sont extraits à l'aide de la *MA Toolbox* de Logan et Slaney [2].

Les vecteurs de chroma de dimension 12 sont régulièrement extraits du signal, à l'aide d'une décomposition en trames de 4096 échantillons chacune et avec une période de 1024 échantillons. L'extraction est faite à l'aide des scripts d'Ellis [3].

2. http ://www.ofai.at/~elias.pampalk/ma/documentation.html
3. http ://labrosa.ee.columbia.edu/projects/coversongs/

Calcul des critères audio Les différentes séquences de descripteurs considérées sont utilisées pour le calcul des critères audio, qui sont exprimés à l'échelle des temps musicaux. Nous utilisons cette échelle d'étude à l'instar d'un ensemble de travaux sur le sujet [MND09, KS10]. Ceci permet de travailler sur une représentation des morceaux robuste aux variations de tempo [Jeh05] : nous considérons ici que ces variations sont généralement liées à l'expressivité des morceaux de musique plutôt qu'à leur structure. On estime les instants associés aux temps d'un morceau à l'aide de l'estimateur d'Ellis [4].

Les critères de rupture d'homogénéité et de détection d'événements sont calculés sur les vecteurs de descripteurs exprimés à l'échelle des trames fixes. Ceci permet de garantir un nombre de vecteurs supérieur à leur dimension dans les différentes fenêtres d'analyse, et d'éviter les problèmes liés à un mauvais conditionnement des matrices de covariance. Les blocs structurels que nous recherchons ont une durée en général comprise entre 10 et 25 s. Nous considérons des fenêtres d'analyse centrées sur chaque temps musical estimé et de durée 12 s, ce qui correspond à trois quarts de la durée typique des blocs structurels. Le fait d'imposer une taille de fenêtre en secondes permet de conserver une échelle d'analyse comparable d'un morceau à l'autre, dans le cas où la période des temps musicaux estimés varie beaucoup. Ainsi, N correspond au nombre de trames de taille fixe contenues dans un intervalle de 6 s. Dans le cas du critère d'événement court, on choisit L égal au nombre de ces trames contenues dans un intervalle de 1 s.

Pour calculer le critère de rupture de répétition, nous avons exprimé les séquences de descripteurs à une résolution temporelle plus grossière. Pour des raisons d'ordre historiques, nous avons sous-échantillonné les séquences d'un facteur quatre. Pour chaque temps estimé nous avons considéré une fenêtre d'analyse centrée sur un temps estimé et couvrant un ensemble de $2N$ vecteurs de descripteurs correspondant à une durée de 12 s, de même que pour les deux autres critères audio [5].

Afin de traiter les effets de bords lors du calcul des ces trois critères, nous avons choisi d'allonger artificiellement les séquences de vecteurs de descripteurs du morceau à ses extrémités. On ajoute respectivement N et $N-1$ vecteurs nuls au début et à la fin de la séquence de descripteurs (*zero-padding*). En pratique, les critères donnent des valeurs très élevées sur les premiers et derniers temps du morceau, qui sont finalement filtrées par le processus de mise en valeur des pics dominants décrit dans le paragraphe suivant. Ainsi ce choix n'a pas d'incidence sur nos résultats.

Pour le calcul du coût associé au critère morphologique, nous avons considéré la description en terme de vecteurs de chroma exprimés à l'échelle des snaps dont l'extraction est explicitée dans la partie 4.2.2. Nous utilisons une fenêtre d'analyse de taille $4N$ avec $N = 4$ vecteurs de descripteurs.

Mise en valeur des pics dominants en vue de leur sélection Les critères audio, calculés sur l'ensemble du morceau, permettent de mettre en valeur un certain nombre d'instants pour lesquels la présence d'une frontière structurelle est vraisemblable, par l'intermédiaire de leurs pics.

Plusieurs méthodes utilisent des seuils adaptatifs afin de détecter ces pics [Got03, MND09, KS10]. Nous nous proposons d'utiliser un critère de décision qui s'est avéré efficace pour la segmentation parole-musique, dans le cadre de travaux menés à l'IRISA

4. http ://labrosa.ee.columbia.edu/projects/coversongs/

5. Bien que cela n'affecte pas nos conclusions, il aurait été plus simple de calculer ce critère directement sur la séquence de vecteurs de descripteurs exprimée à l'échelle des temps musicaux.

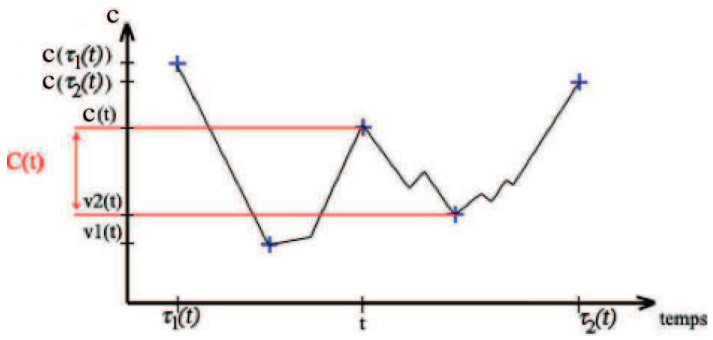

FIGURE 4.5 – Visualisation des quantités utilisées pour le calcul du critère de Seck sur la courbe c à l'instant t.

par Seck *et al.* [SBB00]. Celui-ci identifie et met en valeur les pics dominants d'une courbe relativement aux pics voisins. Nous désignons par la suite cette opération par le terme *filtrage*.

Soit $c(t)$ la courbe à filtrer. Pour tout t, on recherche les premiers instants $\tau_1(t) < t$ et $\tau_2(t) > t$ tels que $c(\tau_1(t)) > c(t)$ et $c(\tau_2(t)) > c(t)$. On calcule ensuite les valeurs $v_1(t) = \min_{\tau_1(t)<i<t} c(i)$ et $v_2(t) = \min_{t<i<\tau_2(t)} c(i)$ puis $u(t) = \max\{v_1(t), v_2(t)\}$. Le critère filtré est obtenu par la formule suivante : $C(t) = c(t) - u(t)$. Tous les points qui ne sont pas des maxima locaux ont ainsi une valeur nulle.

Il s'agit donc de retrancher à l'altitude de chaque pic celle de la vallée la plus haute parmi les deux vallées formées par le pic courant et ses voisins dominants les plus proches (figure 4.5).

Par la suite nous utiliserons systématiquement ce filtrage pour l'étude des différents critères. Soulignons que la combinaison linéaire a lieu après le filtrage, et ce afin d'éviter de perdre certains pics mis en évidence par les critères considérés séparément.

Combinaison des critères On choisit de normaliser chaque critère à combiner en les divisant par leur valeur maximale, ce qui permet de restreindre la grille de valeurs dans laquelle sont recherchés les poids optimaux à $\lambda_i \in [0, 1]$ avec λ_i le poids associé au critère ϕ_i.

4.1.5.2 Estimation de la pulsation structurelle

L'introduction d'une contrainte de régularité nécessite la connaissance de la pulsation structurelle τ. On peut distinguer trois approches pour l'estimation de τ.

τ peut être connue *a priori*, c'est-à-dire estimée par l'intermédiaire de statistiques préalablement réalisées sur un ensemble de morceaux, de manière automatique ou à la main. En particulier, l'annotation des bases présentées dans la partie 3.7 nous a permis de constater qu'un grand nombre de morceaux de musique pop utilise des blocs structurels de taille 16 snaps qui correspondent typiquement à quatre éléments morphologiques de quatre snaps.

τ peut être estimée à partir du morceau analysé. On peut utiliser la périodicité des critères audio s'ils affichent un tel comportement à moyen ou long-terme. Par exemple, l'étude de cette périodicité peut revenir à étudier leur transformée de Fourier, c'est-à-dire localiser les *fréquences structurelles* par la détection des pics du spectre en puis-

sance. Cette recherche doit cependant être limitée à une certaine bande de fréquences, qui doit correspondre à l'échelle d'étude de la structure. En effet, la transformée de Fourier affiche des pics de puissance à la fréquence structurelle associée à τ, mais aussi à ses multiples.

Enfin, τ peut aussi être choisie *a posteriori*. À partir d'un ensemble de pulsations structurelles plausibles établi *a priori*, on calcule les segmentations pour chacune de ces pulsations puis on les classe par pertinence selon un critère particulier, par exemple leur régularité, c'est-à-dire la divergence de la taille des blocs par rapport à τ.

Dans le cadre de cette thèse, nous étudierons en particulier l'approche *a priori* (*cf.* chapitre 6), qui constitue un premier pas à valider avant de considérer des approches plus compliquées à l'avenir. Nous avons néanmoins proposé un système considérant une première approche pour l'estimation de τ lors des évaluations MIREX et Quaero en 2010 (chapitre 5).

4.1.5.3 Estimation des frontières structurelles sous contrainte

L'objectif est maintenant de trouver, pour un morceau donné, la segmentation qui minimise le coût global de segmentation. Une recherche exhaustive dans l'espace contenant toutes les segmentations possibles d'un morceau serait extrêmement coûteuse en temps de calcul. Le principe de programmation dynamique permet de simplifier cette question en considérant qu'un problème peut être résolu de manière optimale en résolvant de manière optimale les problèmes élémentaires qui le constituent [Bel54]. La recherche de la segmentation de moindre coût d'un morceau de musique va ainsi passer par la recherche des segmentations partielles de moindre coût pour des portions de morceau. Ceci est réalisé à l'aide d'un algorithme de Viterbi, dont le principe général a été décrit dans la partie 2.6.

Cependant, le Viterbi classique ne permet pas de prendre en compte des contraintes sur la durée des blocs. Levy et Sandler [LS06] pallient ce problème en s'appuyant sur la notion de modèle de Markov à durée explicite, encore appelé modèle semi-markovien caché (MSMC) [SZ10]. Ce modèle est un MMC classique auquel on ajoute une distribution de probabilité liée à la durée des états. Ainsi, chaque transition entre états implique de choisir une durée particulière à partir de cette distribution explicite. Ils utilisent un algorithme semi-supervisé d'espérance-maximisation afin d'apprendre les paramètres du MSMC dont le nombre d'états est fixé *a priori*, et utilisent un algorithme de Viterbi afin de déterminer la séquence d'états expliquant le mieux la séquence d'observations $\{x_t\}_{1\leq t\leq T}$. Ceci passe par le calcul récursif de la probabilité *a posteriori* de la meilleure séquence d'états se terminant à l'état i à l'instant t :

$$\delta_t(i) = \max_{1\leq d\leq D} \delta^*_{t-d}(i)p(d|i)p(x_{t-d+1},...,x_t|i) \tag{4.15}$$

$$\delta^*_t(j) = \max_{1\leq i\leq M} \delta_t(i)a_{i,j} \tag{4.16}$$

avec D la durée maximale d'un état et M le nombre d'états du MSMC, fixés *a priori*, et $a_{i,j}$ la probabilité de transition de l'état E_i vers l'état E_j.

Notre algorithme correspond à celui de Levy et Sandler mais reformulé dans le cadre présenté dans la partie 4.1.1. Ceci permet d'étendre leur méthode à toute combinaison de critères, la leur étant spécifique à un critère d'homogénéité particulier. On a l'algorithme suivant :

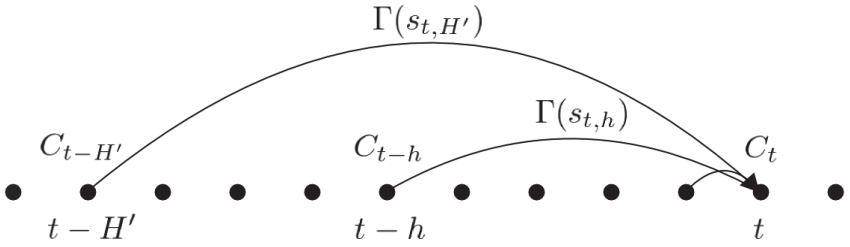

FIGURE 4.6 – Prédécesseurs potentiels pour l'indice temporel t, et leurs coûts.

Soit X un morceau de musique décrit par une séquence de descripteurs $\{x_t\}_{1\leq t\leq T}$. On note $s_{t,h}$ le bloc associé à $X_{t-h}^t = \{x_{t-h}, ..., x_{t-1}\}$, la séquence de descripteurs de taille h qui précède l'instant t. Notons H la taille maximale de cet *historique* de t [6].

- *Initialisation* $(t = 1)$
  On pose $S_1 = \{[0, 1[\}$ et $C_1 = 0$.
- *Pour* $t = 2$ *à* $T - 1$
  Soit $\{t - h\}_{1\leq h\leq H'}$ l'ensemble des prédécesseurs potentiels de l'instant t, avec $H' = \min(t - 1, H)$. On note S_{t-h} la segmentation optimale de la portion de morceau X_1^{t-h} de coût C_{t-h}, comme illustré dans la figure 4.6. La segmentation optimale S_t de X_1^t est la segmentation partielle parmi $\{S_{t-h}\}_{1\leq h\leq H'}$ qui, étendue jusqu'à t, est de coût minimal. Plus précisément, on évalue :

  1. le coût du bloc $s_{t,h}$, noté $\Gamma(s_{t,h})$ pour $1 \leq h \leq H'$,
  2. $b(t) = \text{argmin}_{1\leq h\leq H'}\{C_{t-h} + \Gamma(s_{t,h})\}$, et
  3. $C_t = C_{t-b(t)} + \Gamma(s_{t,b(t)})$.

  on peut noter que $S_t = S_{t-b(t)} \cup s_{t,b(t)}$.

La segmentation optimale pour X, notée S_{opt} et associée au coût C_{N+1}, est obtenue en rassemblant en marche arrière les prédécesseurs optimaux qui ont été rangés dans $b(t)$ (*backtracking*). Les indices temporels associés $\{t_k\}_{1\leq k\leq k_{\text{opt}}}$ sont ensuite retrouvés grâce à la procédure récursive suivante :

1. $t_{k_{\text{opt}}+1} = N + 1 - b(N + 1)$,
2. $t_k = t_{k+1} - b(t_{k+1})$, pour $1 \leq k \leq k_{\text{opt}}$.

k_{opt} correspond au nombre de frontières structurelles de S_{opt} obtenues à l'issue de ce processus de *backtracking*.

4.2 Approche pour l'estimation des étiquettes sémiotiques

Le travail présenté dans cette partie se démarque de ce qui précède par son caractère exploratoire. Nous proposons ici une approche pour l'étiquetage des blocs structurels reposant sur la modélisation du morceau de musique par un automate probabiliste à états finis. On considère ici une description symbolique des morceaux étudiés, ce qui présente l'intérêt d'intégrer des représentations différentes du signal audio comme des partitions, ou des transcriptions issues d'autres algorithmes (accords, mélodie, *etc.*).

6. Typiquement, $H = T$, mais des tailles plus petites peuvent être utilisées, comme des multiples de la pulsation structurelle τ.

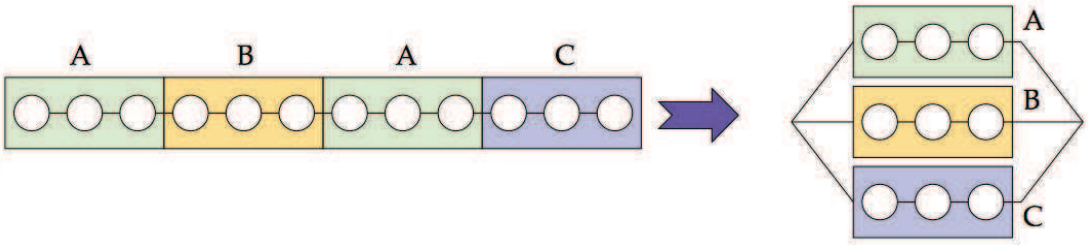

FIGURE 4.7 – Factorisation des séquences d'états semblables en branches d'automate. Si la séquence d'états associée à un bloc structurel contient 16 états ou plus, on ne tient compte que de la séquence d'états associés aux trois premiers quarts du bloc, que l'on suppose correspondre à ses trois premiers éléments morphologiques.

4.2.1 Formulation

Considérons un morceau de musique par le biais d'une séquence de descripteurs exprimée à une échelle particulière, par exemple à l'échelle du snap. Celui-ci peut être modélisé par une séquence d'états d'un automate à états finis, et les descripteurs comme des observations issues de cette séquence d'états.

Nous supposons à ce stade que les frontières structurelles sont connues. L'automate factorise les blocs structurels semblables par leurs séquences d'états en *branches d'automates*. De cette manière, chaque classe de bloc (c'est-à-dire chaque étiquette sémiotique) est représentée par une branche, comme le montre la figure 4.7.

On considère pour chaque automate A_k constitué de k branches la probabilité P_k qu'il produise la séquence d'observations décrivant le morceau de musique. On a :

$$P_k = P_{B_k} P_{O_k} \tag{4.17}$$

- P_{B_k} est la probabilité d'obtenir la séquence de blocs structurels qui représente le morceau, c'est-à-dire la probabilité de choisir la séquence de branches correspondant le mieux aux données observées, et
- P_{O_k} est la probabilité de produire la séquence d'observations X, en ayant choisi la séquence de branches adéquate pour représenter le morceau.

Nous considérons de plus que les tirages des blocs et des observations sont modélisés par des variables aléatoires indépendantes et identiquement distribuées. Notons $P^i_{B_k}$ la probabilité de tirer la branche de l'automate A_k associée au bloc i et $P^i_{O_k}$ la probabilité d'obtenir la séquence d'observations du bloc i sachant que l'on a choisi la branche qui lui est associée. On obtient $P_{B_k} = \prod_i P^i_{B_k}$ et $P_{O_k} = \prod_i P^i_{O_k}$, et on a :

$$P_k = \prod_i P^i_{B_k} \prod_i P^i_{O_k}. \tag{4.18}$$

Explicitons le calcul de P_k par l'exemple suivant : considérons la séquence X segmentée en deux blocs structurels et les deux automates qui la représentent (*cf.* figure 4.8). Dans le cas de l'automate A_2 qui comporte deux branches, nous avons une probabilité de 1 d'obtenir la séquence d'observations associée à un bloc lorsque la branche qui le modélise a été tirée, c'est-à-dire

$$P^1_{O_2} = 1 \text{ et } P^2_{O_2} = 1, \tag{4.19}$$

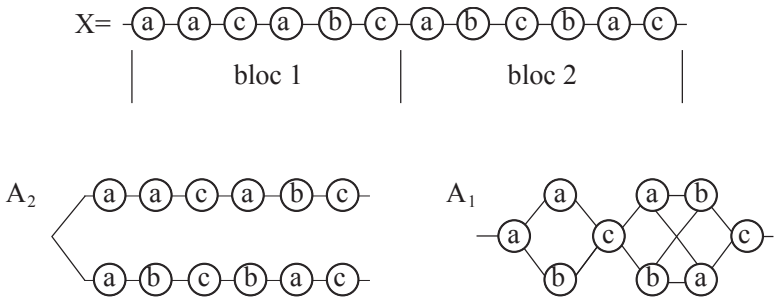

FIGURE 4.8 – Séquence d'observations X constituée de deux blocs structurels et représentée par les automates A_1 et A_2 respectivement constitués de une et deux branches.

et nous avons une probabilité $\frac{1}{2}$ de tirer l'une des deux branches d'où

$$P^1_{B_2} = \frac{1}{2}, P^2_{B_2} = \frac{1}{2}. \tag{4.20}$$

On obtient donc :

$$P_2 = \prod_{i=1}^{2} P^i_{B_2} \prod_{i=1}^{2} P^i_{O_2} = \frac{1}{4}. \tag{4.21}$$

Considérons maintenant l'automate A_1 issu de la fusion des branches associées aux deux blocs. La factorisation des branches fait que, pour chaque branche, on a une probabilité de 1 d'obtenir un a en première position, c en troisième et en dernière position, et une probabilité de $\frac{1}{2}$ d'obtenir la bonne observation pour les trois autres positions. Ainsi

$$P^1_{O_1} = \frac{1}{8} \text{ et } P^2_{O_1} = \frac{1}{8}, \tag{4.22}$$

et l'automate n'étant constitué que d'une branche, on a :

$$P^1_{B_1} = 1, P^2_{B_1} = 1. \tag{4.23}$$

On obtient donc

$$P_1 = P_{B_1} P_{O_1} = \prod_{i=1}^{2} P^i_{B_1} \prod_{i=1}^{2} P^i_{O_1} = \frac{1}{64}. \tag{4.24}$$

Il existe une grande variété d'automates permettant de modéliser un morceau de cette manière. Il nous faut ainsi d'une part définir l'espace des automates pertinents pour notre problème, et d'autre part définir un critère de sélection d'automate.

4.2.1.1 Espace des automates considéré

Soit K le nombre de blocs structurels du morceau de musique. Nous considérons l'ensemble des automates $\{A_k\}_{1 \leq k \leq K}$, k étant le nombre de branches de l'automate A_k. Cet ensemble est construit à partir de l'automate A_K pour lequel chaque branche modélise un bloc différent du morceau par la séquence d'états qui lui est associée. Les autres automates sont obtenus par le regroupement successif de paires de branches, jusqu'à obtenir l'automate A_1 qui ne comprend qu'une seule branche modélisant l'ensemble des blocs de X. Pour tout k, l'automate A_k est obtenu à partir de A_{k+1} de

la manière suivante. On calcule la probabilité P_k associée à tous les automates issus du regroupement de deux branches de l'automate A_{k+1}. A_k correspond à l'automate associé à la probabilité P_k maximale.

4.2.1.2 Utilisation du modèle système - contraste

Le fait que les blocs structurels associés à la même classe d'équivalence (ou étiquette structurelle) partagent des relations paradigmatiques implique que les séquences de descripteurs qui les composent se correspondent au moins en grande partie. D'un point de vue morphologique, et conformément à nos considérations méthodologiques, nous considèrons que deux blocs de même classe sont construits autour du même système porteur, qui est généralement observable sur leurs trois premiers éléments morphologiques (le quatrième pouvant contraster avec le système). Dans cette optique, nous proposons d'utiliser cette conséquence approchée du modèle *système - contraste* dans notre algorithme d'estimation des étiquettes structurelles de manière à associer à chaque bloc structurel les trois quarts de la séquence de descripteurs qui lui est associé si celle-ci est de taille 16 snaps ou plus. On considère que les blocs de taille inférieure correspondent à des blocs réguliers tronqués, dont le quatrième élément morphologique est partiellement réalisé, voir absent. Il est par exemple assez fréquent d'observer des blocs de taille 8 snaps correspondant aux huit premiers snaps d'un bloc de 16 snaps réalisé ailleurs dans le morceau.

Afin d'évaluer la pertinence de l'utilisation de cette conséquence, nous considérerons dans le chapitre 6 deux versions de notre système d'estimation d'étiquettes : l'une associant à chaque bloc la totalité de la séquence de descripteurs (ou d'états) qu'il contient, et l'autre ne leur associant que les trois-quarts de leur séquence si leur taille est supérieure ou égale à 16 snaps.

4.2.1.3 Critères de sélection d'automate

Tous les automates ne vont pas nous intéresser afin de modéliser X. L'automate A_K comportant autant de branches que de blocs structurels peut reproduire avec une plus grande probabilité la séquence X que l'automate A_1 comportant une seule branche mais il comprend un nombre de paramètres plus élevé. En d'autres termes, A_K correspond à une représentation plus précise mais moins économique que A_1. Conformément à la notion de structure sémiotique, nous avons besoin d'un critère de sélection de l'automate optimal, correspondant au meilleur compromis entre précision et économie.

Dans ce cadre, nous considérons deux critères. Le premier est une version pondérée du critère d'information d'Akaike, qui est un critère couramment utilisé dans le cadre de la sélection de modèles. Le second est un critère auto-adaptatif expérimental étudié à titre exploratoire.

Critère d'Information d'Akaike pondéré (AIC) Le critère d'information d'Akaike (ou AIC pour *Akaike Information Criterion* [Aka74]) est un critère de sélection de modèles qui pénalise la quantité d'information de chaque modèle considéré en fonction de son nombre de paramètres. Dans notre cas, il s'agit du nombre de probabilités de transition de l'automate auquel nous appliquons une pondération supplémentaire.

En reprenant les notations de la partie 4.2.1, notre critère AIC est défini par :

$$AIC = a_{AIC} n - \log(P_k) \tag{4.25}$$

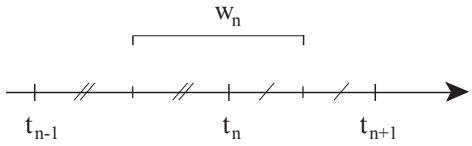

FIGURE 4.9 – Fenêtre d'analyse w_n associée au snap t_n

avec $n \in \mathbb{N}$ le nombre de paramètres du modèle et a_{AIC} un paramètre de pondération qui prend ses valeurs dans $\mathbb{R}^+$.

Critère auto-adaptatif expérimental (CAA) Dans le cadre de la théorie de l'information, la recherche de l'automate de probabilité P_k maximale équivaut à rechercher celui qui possède la quantité d'information $-\log(P_k)$ la plus basse. On observe par ailleurs que pour les morceaux de musique pop, le nombre d'étiquettes structurelles atteint très rarement 1 ou le nombre de blocs structurels annotés. C'est dans cette optique que nous ajoutons une pénalité affine par rapport au nombre de branches permettant de donner la même quantité d'information *a posteriori* à A_K et A_1. Ainsi, l'automate optimal A_k est celui qui minimise le critère suivant :

$$CAA = y_k - \log(P_k) \tag{4.26}$$

où

$$y_k = \frac{\log(P_K) - \log(P_1)}{K-1}(k-1) + \log(P_1). \tag{4.27}$$

4.2.2 Détails d'implémentation

Nous considérons une description symbolique du morceau de musique en terme de vecteurs de chroma quantifiés, et exprimés à l'échelle du snap (décrit au paragraphe 3.4.2). Nous utilisons les vecteurs *Chroma-Pitch* (CP) extraits par la *Chroma Toolbox* de Müller et Ewert [ME11] pour des fenêtres d'analyse contenant 4410 échantillons et espacées de 2205 échantillons. La séquence de vecteurs CP est exprimée à l'échelle des snaps en associant à chaque snap estimé le vecteur issu de la moyenne des vecteurs CP contenus dans la fenêtre d'analyse représentée dans la figure 4.9. La quantification est réalisée à l'aide d'un algorithme de quantification vectorielle (méthode LBG ou Lloyd généralisé, initialisée par division récursive des données selon le barycentre ou *splitting* [Gra84]) pour lequel on a fixé empiriquement à 16 le nombre de classes de vecteurs de chroma.

L'échelle des snaps est obtenue à l'aide des estimateurs des temps musicaux et du premier temps des mesures musicales de Davies ([Dav07, ch. 6], et [SDP09]). Elle est construite en sélectionnant l'échelle des temps musicaux dont la période est la plus proche de 1 seconde et synchronisée sur les premiers temps des mesures.

4.3 Résumé du chapitre

Ce chapitre introduit plusieurs approches pour estimer la position des frontières des blocs structurels, puis les étiquettes sémiotiques de ces blocs.

Nous formulons tout d'abord la question de l'estimation des frontières par l'optimisation d'un coût de segmentation qui fait clairement apparaître les contributions

associées aux critères audio et celles liées aux contraintes structurelles. Nous proposons un ensemble de critères audio, nouveaux et anciens, formulés dans un même cadre en vue de leur combinaison. Il s'agit d'un critère de rupture d'homogénéité, d'un critère de rupture de répétition et d'un critère de détection d'événements localisés. Ceci permet de considérer une approche multicritère dans le but d'étudier la variété des manifestations de la structure sémiotique. Un premier critère morphologique, basé sur le modèle système - contraste, est introduit à titre exploratoire. Nous proposons ensuite une contrainte de régularité structurelle, permettant d'introduire l'hypothèse d'une pulsation structurelle lors de l'estimation des frontières.

Nous nous intéressons enfin à un système d'estimation des étiquettes structurelles basé sur la sélection d'automates probabilistes à états finis prenant en compte certains aspects du modèle système - contraste.

Ces approches sont étudiées par l'intermédiaire de l'évaluation de plusieurs systèmes les mettant en oeuvre dans le chapitre 5 et du diagnostic de leurs implémentations séparées dans le chapitre 6.

Chapitre 5

Évaluation de systèmes d'estimation de la structure sémiotique

Les principaux avancements de cette thèse sont liés à nos participations aux campagnes d'évaluation MIREX et Quaero, qui se déroulent quasiment simultanément chaque année depuis 2009. Celles-ci furent l'occasion d'élaborer un ensemble de systèmes complets et opérationnels en un temps limité, et ainsi de comparer nos approches pour l'estimation de la structure à celles de l'état de l'art. Cette démarche exploratoire implique que les systèmes peuvent ne pas avoir été soumis dans leur configuration optimale. Pour pallier ce problème, nous effectuons un diagnostic des principaux modules constituant nos systèmes dans le chapitre 6.

Les systèmes que nous avons soumis sont influencés par la notion de structure sémiotique, qui a évolué jusqu'à atteindre la spécification présentée au chapitre 3. Ils sont ainsi notamment composés d'un ou plusieurs critères audio et d'une contrainte de régularité pour l'estimation des frontières structurelles. L'organisation globale des différents systèmes est décrite dans la figure 5.1.

Nous présentons et discutons dans ce chapitre les résultats des campagnes MIREX (*structural segmentation task*) et Quaero (tâche 6.5 : *music structuring and summarization evaluation*) de 2010, 2011 et 2012 auxquelles nous avons participé. La première partie nous permet d'introduire le contexte expérimental des différentes campagnes d'évaluations considérées. Chacune de nos participations fait ensuite l'objet d'une par-

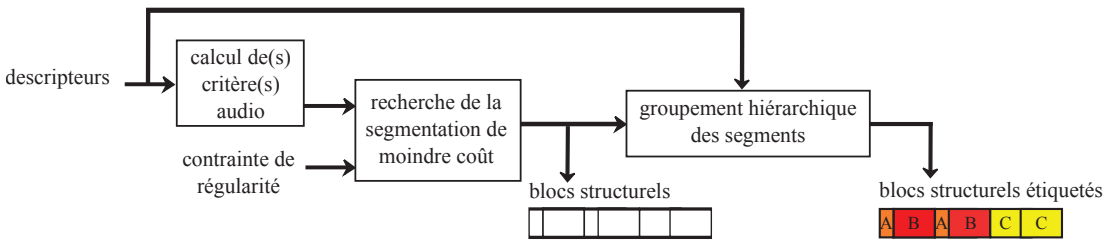

FIGURE 5.1 – Organisation globale des systèmes IRISA10,11 et 12 soumis aux campagnes d'évaluation MIREX et Quaero de 2010, 2011 et 2012. Les descripteurs sont extraits à partir de l'audio à l'aide de *toolboxes* mises à disposition par la communauté MIR.

tie de ce chapitre : nous y présentons le système soumis, ceux des différents participants puis les performances obtenues. La dernière partie présente un bilan des systèmes que nous avons soumis pour l'estimation des frontières et propose d'explorer leur complémentarité par l'évaluation d'un système de fusion des frontières estimées.

5.1 Contexte expérimental

5.1.1 Bases de morceaux utilisées lors des campagnes MIREX et Quaero de 2010 à 2012

Le tableau 5.1 répertorie l'ensemble des bases de morceaux considérées lors des différentes campagnes d'évaluation auxquelles nous avons participé. Tous les morceaux considérés sont au format mono.

La base MIREX09 correspond au rassemblement de plusieurs bases existantes, annotées par Jouni Paulus (Tampere University of Technology TUT), Ewald Peiszer (Vienna University of Technology VUT) et par le laboratoire C4DM (Queen Mary University of London QMUL). La base contient 297 morceaux dont la liste n'est pas disponible actuellement, mais "une grande partie de celle-ci est constituée de morceaux des Beatles" [EBD+11]. Les méthodologies d'annotation de la structure peuvent ainsi varier selon les morceaux et ne sont pas accessibles aujourd'hui. Une description des trois bases composant MIREX09 est disponible dans [PD09].

La base MIREX10 considère l'ensemble des morceau de la base RWC Pop. Cet ensemble est associé d'une part aux annotations obtenues avec la méthodologie de la partie 3.7 (IRISA) et d'autre part à ses annotations originales (AIST) [1]. Il faut noter que les annotations IRISA utilisées dans les campagnes d'évaluation de 2010 à 2012 ont été produites en 2010. Elles ont depuis fait l'objet de révisions, mineures, à l'issue de l'affinement de notre méthodologie d'annotation.

La base du projet SALAMI est originalement constituée de 1383 morceaux issus de divers styles musicaux (pop, jazz, world, classique) et contient un ensemble de morceaux interprétés en concert. Ces morceaux ont été annotés par le CIRMMT (Université Mc Gill de Montréal) en terme de frontières et d'étiquettes structurelles [SBF+11]. Quatre points de vue sont considérés, ce qui implique qu'un fichier d'annotation contient quatre niveaux d'annotation : celui de l'instrument qui prédomine à l'écoute, celui de la fonction (introduction, couplet, refrain...) et celui de la similarité à long et moyen terme [2]. Mentionnons que chaque morceau est annoté par deux annotateurs musicologues : lorsque les annotations produites diffèrent pour un même morceau, elles sont toutes deux conservées et considérées comme deux annotations plausibles. La base considérée dans le cadre de MIREX et que l'on note MIREX12 par la suite est constituée de 1000 morceaux choisis au hasard parmi les 1383 évoqués plus haut. Les morceaux doublement annotés se voient arbitrairement attribués l'une des deux annotations disponibles. MIREX a choisi d'associer à chaque morceau le niveau d'annotation correspondant au point de vue de la similarité à long terme.

Les bases associées à Quaero sont constituées de morceaux issus de divers styles musicaux (pop, rock, électro...) choisis par l'IRCAM et correspondent à la description de la partie 3.7. Ces morceaux ont été annotés selon la méthodologie du chapitre 3

1. Ce dernier ensemble d'annotation a été produit par l'AIST et est disponible à l'adresse `http://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/`.

2. `http://www.music.mcgill.ca/~jordan/salami/SALAMI-Annotator-Guide.pdf`

base de morceaux	taille	frontières	étiquettes	annotateurs
MIREX09	297	oui	oui	TUT, UTV, QMUL
MIREX10 (IRISA) Dev+Test	50+50	oui	non	IRISA
MIREX10 (AIST)	100	oui	oui	AIST
MIREX12	1000	oui	oui	Université Mc Gill Montréal
Quaero Dev09	20	oui	oui	IRCAM
Quaero Test09+10 (IRISA)	49+45	oui	non	IRISA
Quaero Test09+10 (IRCAM)	49+45	oui	oui	IRCAM
Quaero Test11 (IRISA)	45	oui	non	IRISA
Quaero Test11 (IRCAM)	45	oui	oui	IRCAM
Quaero Test12	40	oui	oui	IRCAM
Eurovision Dev+Test	61+63	oui	non	IRISA

TABLEAU 5.1 – Inventaire des bases de morceaux utilisées lors des campagnes d'évaluation MIREX et Quaero de 2010 à 2012. La taille des bases correspond à leur nombre de morceaux.

(IRISA) et par l'IRCAM. Celles de l'IRCAM résultent de la fusion des annotations suivant les points de vue définis dans [PD09] et présentés dans la partie 2.1.2 en une seule annotation mono-dimensionnelle. Ainsi, deux portions ayant les mêmes étiquettes selon les mêmes points de vue se voient attribués la même étiquette mono-dimensionnelle, et des étiquettes mono-dimensionnelles différentes dans le cas contraire. L'ensemble du corpus Quaero comprend une base de développement (Dev09) et trois bases de test (Test09+10 pour 2010, Test11 pour 2011 et Test12 pour 2012).

La base Eurovision correspond à celle décrite dans la partie 3.7. On notera (Dev) la partie d'Eurovision utilisée pour le développement des algorithmes et (Test) la partie pour leur test. Cette séparation (61 morceaux pour Dev et 63 pour Test) a été effectuée par l'IRIT, en charge de la production des résultats dans le cadre de la campagne Quaero.

Précisons que les annotations des bases MIREX10 (IRISA), Eurovision et Quaero (IRISA) ne concernent que les frontières structurelles et ne contiennent donc pas d'étiquettes structurelles. Seules les métriques d'évaluation de l'estimation des frontières ont ainsi pu être calculées sur ces bases. Ceci justifie la mention de symboles "-" apparaissant dans les prochains tableaux de résultats pour les métriques concernant l'évaluation de la structure complète (frontières et étiquettes). Ces étiquettes n'existaient pas à l'époque de ces évaluations car la méthodologie d'annotation était focalisée sur l'annotation des frontières structurelles. Ces bases sont aujourd'hui complètes et disponibles en ligne [3].

Le tableau 5.2 décrit quelles bases ont été utilisées pour développer (Dev) et tester (Test) les algorithmes que nous avons soumis aux campagnes d'évaluation MIREX et Quaero entre 2010 et 2012. La base de test de la campagne Quaero en 2010 est constituée des bases Test09 et Test10, que l'on abrège par Test09+10. Les bases MIREX10 et Eurovision ont été divisées en deux parties comparables par l'IRIT dans le cadre des campagnes d'évaluation Quaero.

3. `http://musicdata.gforge.inria.fr/structureAnnotation.html`

campagne	Dev	Test
MIREX 2010	Quaero Dev09	MIREX09, MIREX10 (IRISA)
MIREX 2011	MIREX10 (IRISA)	MIREX09, MIREX10 (IRISA)
MIREX 2012	MIREX10 (IRISA)	MIREX09, MIREX10 (IRISA), MIREX10 (AIST), MIREX12
Quaero 2010	Quaero Dev09	MIREX10 (IRISA), Quaero Test10 (IRISA et IRCAM)
Quaero 2011	MIREX10 (IRISA)	MIREX10 (IRISA, Test), Eurovision (Test), Quaero Test10 et Test11 (IRISA et IRCAM)
Quaero 2012	MIREX10 (IRISA, Dev), Eurovision (Dev)	Eurovision (Test), Quaero Test10, Test11 et Test12 (IRISA et IRCAM)

TABLEAU 5.2 – Bases de morceaux utilisées lors des campagnes d'évaluation MIREX et Quaero, pour chaque année, de 2010 à 2012. "Dev" et "Test" correspondent respectivement aux bases de développement et de test des algorithmes soumis.

5.1.2 Métriques d'évaluation

Il existe un grand nombre de métriques utilisées pour l'évaluation des systèmes d'estimation de la structure. Certaines n'évaluent que la position des frontières structurelles estimées par rapport aux frontières de référence, d'autres comparent la structure estimée et celle de référence dans leur ensemble (c'est-à-dire les frontières et les étiquettes structurelles).

Nous portons notre attention sur la F-mesure et le *boundary hit rate* F_{br} pour les frontières, sur la F-mesure dyadique et le score de modélisation pour la structure complète. Ces métriques sont conceptuellement simples, ont l'avantage d'être symétriques et sont utilisées soit dans la campagne de MIREX soit celle de Quaero [4].

5.1.2.1 Évaluation de la segmentation

F-mesure, précision et rappel L'évaluation d'un système de segmentation consiste à comparer les frontières estimées par ce système aux frontières de référence pour un ensemble de morceaux donné. Cette comparaison est réalisée par l'intermédiaire d'un ensemble de métriques d'évaluation fréquemment utilisées dans le cadre de la recherche automatique d'information (*Information Retrieval*) [vR79] : la précision, le rappel et la F-mesure. Nous allons les définir dans le cadre de l'estimation des frontières structurelles.

Soient f_R l'ensemble des frontières de référence d'un morceau de musique, et f_E l'ensemble des frontières estimées par le système évalué, on définit la précision (P) et le rappel (R) comme suit :

$$P = \frac{|f_E \cap f_R|}{|f_E|}, \tag{5.1}$$

$$R = \frac{|f_E \cap f_R|}{|f_R|}, \tag{5.2}$$

4. Ces deux campagnes d'évaluation ont malheureusement peu de métriques en commun. Notons cependant que la F-mesure et le *boundary hit rate* F_{br} sont deux versions de la même métrique.

où $|X|$ désigne le nombre d'éléments de l'ensemble X. La précision correspond ainsi au taux de frontières estimées correspondant à des frontières de référence, et le rappel au taux de frontières de référence correspondant aux frontières estimées. La correspondance est approchée : on considère que deux frontières correspondent si leur écart temporel est inférieur à une certaine valeur de tolérance, typiquement 0.5 s [TLPG07] ou 3 s [LS08]. Ces valeurs sont par exemple utilisées dans de cadre de la campagne d'évaluation MIREX [5]. Notons qu'une même frontière peut être associée à plusieurs autres lors du calcul de la précision et du rappel, contrairement aux *boundary hit rates* définies plus loin.

La F-mesure est la moyenne harmonique de ces deux quantités :

$$F = \frac{2PR}{P+R} \tag{5.3}$$

Cette quantité se détériore lorsque P ou R diminue ; en particulier elle suit le comportement du plus faible des deux dans le cas où l'un est négligeable devant l'autre (par exemple $F = 2R$ si $R << P$).

Boundary hit rates La définition des *boundary hit rates* P_{br}, R_{br} et F_{br} est respectivement la même que celle de la précision, du rappel et de la F-mesure [TLPG07]. Il s'agit cependant ici de considérer que chaque frontière estimée (respectivement de référence) ne peut être associée qu'à une seule frontière de référence (respectivement estimée).

5.1.2.2 Évaluation de la structure complète

Plusieurs métriques ont été proposées afin de comparer la structure de référence aux estimations produites par les algorithmes.

Mesures de précision, rappel et F-mesure pour les groupements dyadiques de trames [LS08] Le calcul des mesures de précision, rappel et F-mesure pour les groupements dyadiques de trames (*pairwise precision, recall and F-measure*, que l'on note respectivement pP, pR, pF par la suite) reprend les formules de celles utilisées dans le cas des frontières structurelles mais compare des objets différents. On suppose ici que le morceau est divisé en petites fenêtres successives qui ne se recouvrent pas (en pratique il s'agit de trames de 100 ms). Il s'agit de comparer les paires de fenêtres associées à la même étiquette structurelle par l'algorithme à celles associées à la même étiquette de référence. Soient p_E l'ensemble des paires de fenêtres associées à la même étiquette estimée et p_R l'ensemble des paires de fenêtres associées à la même étiquette de référence, on a :

$$pP = \frac{|p_E \cap p_R|}{|p_E|}, \tag{5.4}$$

$$pR = \frac{|p_E \cap p_R|}{|p_R|}, \tag{5.5}$$

et

$$pF = \frac{2pPpR}{pP+pR}. \tag{5.6}$$

5. `http://www.music-ir.org/mirex/wiki/2011:Structural_Segmentation`

Score de modélisation [Pee07] On considère pour chaque classe de segments structurels de l'annotation de référence la classe de segments structurels estimée dont les segments recouvrent le mieux les segments qu'il contient. Le score de modélisation correspond au taux de recouvrement des segments des classes de référence par les classes estimées une fois ces associations faites.

Plus précisément, soit r_i le vecteur temporel tel que $r_i(t)$ prend la valeur 1 lorsque la classe de segments structurels de référence i existe à l'instant t et 0 sinon. On définit de la même manière le vecteur e_j pour la classe de segments structurels estimée j. On associe à chaque classe de référence i la classe estimée $k(i)$ maximisant le produit scalaire $\langle e_{k(i)}|r_i\rangle$.

Le score de modélisation (ou *modeling score*) Σ_{mod} correspond à la somme des produits scalaires obtenus normalisé par la durée du morceau :

$$\Sigma_{\text{mod}} = \frac{\sum_i \langle e_{k(i)}|r_i\rangle}{\sum_i \sum_t r_{k(i)}(t)} \tag{5.7}$$

Ce score tend vers 1 lorsque les structures estimée et de référence sont les mêmes. Il tend vers 0 lorsqu'elles diffèrent.

5.2 Campagnes d'évaluation MIREX et Quaero de 2010

5.2.1 Motivation et description de l'algorithme proposé

Au début de la thèse, nous nous sommes focalisés sur l'annotation manuelle des frontières structurelles de plusieurs bases de morceaux (RWC Pop, Quaero), ce qui a permis de mettre en avant la variété des indices structurants observés selon leurs strates musicales et la manière dont ils se manifestent sur ces strates, ainsi que d'émettre une première hypothèse de régularité structurelle.

Il en a résulté le développement d'un algorithme d'estimation de la structure en deux parties : une première partie relative à l'estimation des frontières structurelles et une seconde partie relative à l'étiquetage des blocs structurels.

L'algorithme analyse les morceaux de musique par l'intermédiaire de deux séquences de descripteurs numériques exprimés à l'échelle des temps musicaux : les descripteurs MFCCs et les vecteurs de chroma. Leur processus d'extraction est décrit dans la partie 4.1.5.1.

La première partie de l'algorithme combine plusieurs composantes : trois critères audio et une contrainte de régularité. Les critères utilisés sont le critère de rupture d'homogénéité et de détection d'événements calculés sur les descripteurs MFCCs, et le critère de rupture de répétition calculé sur les vecteurs de chroma décrits dans la partie 4.1.2.2. Ils sont filtrés par le critère de Seck, normalisés[6] puis sommés. La contrainte de régularité utilisée a été choisie empiriquement dans le cadre de cette campagne. Il s'agissait de pénaliser davantage les segments de taille inférieure à la pulsation structurelle τ par rapport aux segments de taille supérieure à celle-ci. Nous avons utilisé la fonction convexe

$$\Psi_0(s) = \frac{m}{\tau} + \frac{\tau}{m} - 2 \tag{5.8}$$

6. À chaque pic de critère est associé son image par la fonction d'erreur complémentaire gaussienne, ce qui est un moyen de ramener les distributions de pics des critères à des amplitudes comparables.

pour s un segment de taille m non nul. Ces différentes composantes sont combinées pour former le coût de segmentation des équations 4.1 et 4.2, où $\Phi(s_k)$ correspond à la somme des valeurs prises par la courbe résultant de la combinaison des critères entre les instants de début et de fin de s_k et Ψ correspond à contrainte de régularité mentionnée ci-dessus. Le paramètre de pondération λ est empiriquement fixé à 0.33. τ est estimée en sélectionnant la durée associée à la valeur maximale du cepstre calculé sur le critère de rupture d'homogénéité lui-même filtré par le critère de Seck, autour d'une durée cible déduite de la durée du morceau [7]. Nous utilisons ce critère car il s'est avéré avoir le comportement le plus régulier parmi les trois critères implémentés sur un sous-ensemble de la base d'étude MIREX10 (IRISA). La recherche de la meilleure segmentation passe par l'optimisation de ce coût de segmentation à l'aide de l'algorithme de Viterbi décrit dans la partie 4.1.5.3.

La seconde partie de l'algorithme est issue d'un système d'estimation développé au début de la thèse. Elle consiste à modéliser chaque bloc structurel estimé par une distribution gaussienne issue de sa séquence de descripteurs MFCC. Chaque classe de blocs est modélisé par la distribution gaussienne issue des descripteurs des blocs qu'elle contient. La distance entre deux blocs ou deux classes de blocs est évaluée à l'aide d'une *mesure de vraisemblance gaussienne symétrisée* entre leurs modèles gaussiens définie dans [BMM95]. Celle-ci s'apparente à une version symétrique de la divergence de Kullback-Leibler. Les blocs structurels sont regroupés à l'aide de l'algorithme de regroupement hiérarchique suivant. Initialement, chaque bloc est caractérisé par une gaussienne et est contenu dans une classe différente. A chaque itération, on calcule la distance entre toutes les paires de classes. On fusionne les deux classes de distance minimale. La sélection du nombre d'itérations est effectuée *a posteriori* en modélisant les distances minimales par une distribution bi-gaussienne. La première gaussienne modélise les faibles distances entre les blocs ou classes de blocs de même étiquette structurelle, et la seconde modélise les distances entre les blocs ou classes de blocs d'étiquettes structurelles différentes. La sélection des distances associées à chaque gaussienne est effectuée à l'aide d'un algorithme des K-moyennes pour $K = 2$. Le nombre de distances associé à la première gaussienne détermine le nombre d'itérations n_{iter} du groupement hiérarchique, qui est enfin ajusté par la relation linéaire

$$n_{\text{iter}'} = a n_{\text{iter}} + b \tag{5.9}$$

où a et $b \in \mathbb{R}$.

Le lecteur pourra se référer à [SBV10b] pour plus d'information sur cette dernière partie de l'algorithme, qui n'est pas au centre de la thèse.

Deux versions de cet algorithme ont été soumises. Elles ne diffèrent que par la valeur des paramètres d'ajustement du nombre d'itérations de l'algorithme de groupement hiérarchique : IRISA10_2 correspond au paramétrage ($a = 0.5766, b = 0.3522$) et IRISA10_2 à ($a = 1, b = 0$). Le premier couple de paramètres a été réglé sur la base Quaero Dev09. Le second paramétrage revient à ignorer l'étape d'ajustement de n_{iter}.

7. Cette durée cible $\tau_0 = \sqrt{T}$ pour un morceau de durée T, est obtenu par la minimisation du *contexte informatif prédominant* défini dans [BLSV10a]. Il s'agit d'un *a priori* sur le nombre de trames utiles pour prédire le contenu acoustique d'une trame particulière pour un morceau de musique donné.

5.2.2 Participants

L'algorithme GP7 [Pee10] utilise quatre séries de descripteurs : trois de type timbral (MFCCs et ses moments) et un de type tonal (vecteurs de chroma). Ces descripteurs sont combinés par la somme pondérée de leurs matrices de similarité. L'estimation des frontières structurelles passe par la segmentation de la matrice résultante selon un critère d'homogénéité (par une méthode proche de celle de la fonction de nouveauté décrite dans la partie 2.6), et l'étiquetage des segments est effectué selon un critère de répétition [Pee07].

Les algorithmes MND1, MHRAF2 et WB1 fondent tous les trois leur estimation de la structure sur l'analyse des répétitions dans la séquence de vecteurs de chroma des morceaux de musique, en tenant compte de contraintes sur la durée minimale des segments structurels visés.

MHRAF2 effectue une décomposition hiérarchique des morceaux [MHRF11]. Il s'agit de détecter les séquences de vecteurs de chroma les plus longues et les plus semblables selon une technique de déformation temporelle dynamique parente de celle décrite dans la partie 2.6 à l'échelle du morceau. Celle-ci est de plus robuste aux transpositions. Chaque itération de l'algorithme détecte des répétitions de plus en plus courtes. Ceci permet de construire une représentation en arbre des différentes répétitions. La structure estimée par l'algorithme correspond à un niveau particulier de cet arbre : les segments structurels répétés ont la même étiquette, et les segments non référencés ont des étiquettes différentes.

MND1 calcule une matrice de similarité dérivée des vecteurs de chroma, la filtre et localise les répétitions en détectant les diagonales de forte similarité. Cette détection est effectuée à l'aide d'un seuil adaptatif. Les répétitions sont contraintes de débuter sur le premier temps d'une mesure musicale ainsi que d'avoir une durée multiple de quatre temps musicaux. Enfin le morceau est décomposé en privilégiant les segments correspondant aux répétitions les plus semblables et les plus longues [MND09].

WB1 utilise une technique de décomposition matricielle sous contrainte de parcimonie (SI-PLCA) afin d'estimer un ensemble limité de séquences de vecteurs de chroma de taille fixée (70 temps musicaux) constituant une base sur laquelle décomposer la séquence de vecteurs de chroma du morceau de musique. Il s'agit donc d'estimer les segments prototypiques avant de segmenter le morceau selon eux. La segmentation est réalisée en attribuant à chaque instant le prototype le plus vraisemblable à l'aide d'un algorithme de Viterbi [WB10].

5.2.3 Résultats obtenus

Résultats concernant l'estimation des frontières structurelles Le tableau 5.3 répertorie les performances liées à l'évaluation des frontières structurelles estimées par les algorithmes soumis lors de la campagne d'évaluation MIREX 2010. Il s'agit des valeurs moyennes des *boundary hit rates* F_{br}, P_{br} et R_{br} sur l'ensemble des morceaux des bases MIREX09 et MIREX10 (IRISA).

On peut noter que les performances moyennes des différents algorithmes sont globalement plus élevées sur MIREX10 (IRISA) que sur MIREX09. Ceci peut venir du fait que les morceaux diffèrent par leur style ou leur contexte de production comme décrit dans la partie 5.1.1, ce qui peut avoir une influence sur la régularité de leur structure. Enfin, les approches méthodologiques utilisées pour ces deux bases diffèrent.

L'approche utilisée pour annoter MIREX10 (IRISA) est cohérente pour son ensemble tandis que MIREX09 rassemble plusieurs bases d'annotateurs différents.

Les valeurs moyennes de F_{br} des algorithmes IRISA10_1 et IRISA10_2, que nous avons soumis, sont de l'ordre de celles des algorithmes de l'état de l'art pour les tolérances de 0.5 s et 3 s pour les deux bases considérées. R_{br} est nettement supérieur à P_{br} pour les deux tolérances pour la base MIREX09, ce qui implique que l'algorithme tend à sur-segmenter les morceaux de musique de cette base. Notons que cette tendance n'est pas observée dans le cas de la base MIREX10 (IRISA) [8]. Il est cependant surprenant d'observer une variation entre les performances de IRISA10_1 et IRISA10_2 pour l'estimation des frontières et dans le cadre de MIREX car leurs algorithmes ne diffèrent pas pour cette partie. Les estimations des deux systèmes coïncident parfaitement pour 94% de la base MIREX10, comme nous le verrons dans la partie 5.6.

On remarque que l'algorithme MND1 se démarque par ses bonnes performances au niveau de l'estimation des frontières structurelles. Cependant, les algorithmes MHRAF2 et WB1, qui reposent sur l'utilisation des critères de répétition au niveau du contenu tonal, produisent des F_{br} semblables à GP7 qui base sa segmentation sur un critère d'homogénéité au niveau timbral. L'algorithme d'estimation des frontières de MND1 reposant lui aussi sur un critère de répétition au niveau tonal, il semble qu'il soit avantagé d'une part par ses contraintes sur le début et sur la longueur des segments, et d'autre part par la stratégie de sélection des répétitions les plus pertinentes aboutissant à la segmentation du morceau.

Le tableau 5.4 répertorie les performances liées à l'évaluation des frontières structurelles estimées par les algorithmes soumis lors de la campagne d'évaluation Quaero 2010. Les métriques utilisées sont légèrement différentes (précision, rappel et F-mesure) et sont considérées avec une tolérance intermédiaire : tol=2 s. On remarque que les résultats des algorithmes IRISA10_1 et IRISA10_2 sont plus faibles en comparaison de l'algorithme GP7 sur les bases Quaero Test09+10 (IRISA et IRCAM), ce qui peut s'expliquer par le fait qu'ils soient constitués de morceaux de styles plus hétérogènes que MIREX10 (IRISA). IRISA10_1 et IRISA10_2 ont tendance à sur-segmenter les morceaux ($P < R$), ce qui peut être causé par l'estimation d'une pulsation structurelle trop petite et l'estimation d'une structure trop régulière en comparaison de la complexité structurelle de certains morceaux de musique de la base.

Résultats concernant l'estimation de la structure complète Le tableau 5.3 rassemble des performances des différents algorithmes soumis pour la base MIREX09 car la base MIREX10 (IRISA) n'était alors annotée qu'en terme de frontières structurelles. On observe que les performances de IRISA10_1 sont supérieures à celles de IRISA10_2 ce qui montre l'utilité de l'ajustement du nombre d'étapes du groupement hiérarchique pour l'étiquetage à l'aide des paramètres d'ajustement (a, b) [9]. Les mesures pF obtenues sont cependant un peu plus faibles que celles des autres participants, ce qui est en faveur de l'utilisation d'un critère de répétition devant un critère d'homogénéité pour l'étiquetage des segments structurels (cette tendance ne sera pas cependant pas observée lors de la campagne d'évaluation de 2011). On notera l'efficacité des algorithmes d'étiquetage de MHRAF2 et WB1 qui, malgré des F_{br} en retrait au vu des

8. Rappelons que le seul paramètre intervenant dans notre algorithme de segmentation, λ, a été réglé empiriquement sur un petit nombre de morceaux de MIREX10 (IRISA).

9. Ceci peut être aussi influencé par la différence sur les estimations des frontières structurelles observée dans ce même tableau (*boundary hit rates*).

Base MIREX09									
	segmentation (tol = 0.5s)			segmentation (tol = 3s)			étiquetage		
Participants	F_{br}	P_{br}	R_{br}	F_{br}	P_{br}	R_{br}	pF	pP	pR
IRISA10_1	**21.72**	**18.06**	**29.18**	**56.67**	**47.03**	**75.70**	**50.16**	**59.79**	**47.32**
IRISA10_2	**21.93**	**18.97**	**29.28**	**55.93**	**47.29**	**74.47**	**49.28**	**41.81**	**73.40**
GP7	18.13	14.38	25.70	50.14	40.04	70.31	53.59	63.06	50.59
MHRAF2	18.53	20.17	18.07	50.79	54.86	49.88	55.46	50.51	76.86
MND1	**32.46**	33.51	33.40	**60.74**	62.57	62.63	**61.26**	55.41	74.35
WB1	20.04	19.58	21.80	47.53	46.29	51.62	54.40	53.02	63.62
Base MIREX10 (IRISA)									
	segmentation (tol = 0.5s)			segmentation (tol = 3s)			étiquetage		
Participants	F_{br}	P_{br}	R_{br}	F_{br}	P_{br}	R_{br}	pF	pP	pR
IRISA10_1	**23.38**	**24.33**	**23.71**	**61.01**	**61.90**	**62.16**	-	-	-
IRISA10_2	**23.53**	**25.21**	**23.72**	**60.60**	**62.20**	**61.60**	-	-	-
GP7	22.77	23.25	23.26	57.08	57.52	59.07	-	-	-
MHRAF2	20.30	31.63	15.31	48.64	75.19	36.90	-	-	-
MND1	**35.89**	44.08	32.27	60.51	73.64	54.39	-	-	-
WB1	29.05	36.18	24.88	58.19	72.00	50.04	-	-	-

TABLEAU 5.3 – Résultats de l'évaluation de l'estimation des frontières structurelles et de la structure complète obtenues dans le cadre de la campagne MIREX 2010 sur les bases MIREX09 et MIREX10 (IRISA). Les *boundary hit rates* F_{br}, P_{br} et R_{br} sont calculées pour les tolérances (tol) de 0.5 s et 3 s.

Base Quaero Test09+10 (IRISA)				
Participants	F	P	R	score de modélisation
IRISA10_1	**38.53**	**36.48**	**43.54**	-
IRISA10_2	**38.53**	**36.48**	**43.54**	-
GP7	**41**	41	42	-
Base Quaero Test09+10 (IRCAM)				
Participants	F	P	R	score de modélisation
IRISA10_1	**46.34**	**42.65**	**54.29**	**55.45**
IRISA10_2	**46.34**	**42.65**	**54.29**	**48.97**
GP7	**53**	54	56	**62**

TABLEAU 5.4 – Résultats de l'évaluation Quaero 2010 sur les bases Quaero Test09+10 (IRISA et IRCAM). L'ensemble des métriques sont exprimées en %. Les F, P et R ont été calculés pour une tolérance de 2 s.

autres systèmes, permettent d'obtenir les deuxièmes et troisièmes places du classement selon les pF. Ceci peut venir du fait que les étiquettes des segments structurels sont estimées conjointement à la segmentation du morceau de musique, par l'estimation des "prototypes" pour WB1 et par la recherche multi-échelles des répétitions,d'une échelle grossière à une échelle fin, pour MHRAF2.

Notons que le classement des algorithmes IRISA10_1, IRISA10_2 et GP7 est conservé si l'on compare les F-mesures dyadiques de MIREX aux scores de modélisation de l'évaluation Quaero de 2010 (tableau 5.4) sur la base Test09+10. La position de GP7 peut être expliquée par le critère de répétition qu'il utilise pour l'étiquetage.

5.3 Campagnes d'évaluation MIREX et Quaero de 2011

5.3.1 Motivation et description de l'algorithme proposé

L'algorithme d'estimation de structure IRISA11 que nous avons soumis aux campagnes MIREX et Quaero en 2011 est constitué de deux parties : une estimation des frontières structurelles qui utilise un critère de répétition et une contrainte de régularité, et une estimation des étiquettes structurelles qui est réalisée par le biais de la sélection adaptative d'une modélisation par des automates à états finis du morceau de musique.

Ces deux estimations se fondent sur une description des morceaux de musique en terme d'accords exprimés à l'échelle du snap. Le choix d'un descripteur de type symbolique est motivé par la volonté d'utiliser des descriptions issues d'autres algorithmes du MIR dans le cadre de l'estimation de structure à partir de l'audio : à notre connaissance, cet axe de recherche n'a pas encore été exploré. Notre collaboration avec l'Université de Tokyo (Sagayama/Ono Laboratory) dans le cadre du projet VERSAMUS [10] nous a notamment permis de travailler avec l'algorithme d'estimation d'accords de Ueda [UUN+10] intégré à l'algorithme soumis. Les accords de type majeur, mineur, augmentés, diminués et de septième sont considérés. On associe enfin à chaque accord un symbole différent afin d'obtenir la séquence de symboles traitée. On considère donc ici que deux accords sont soit identiques, soit n'ont rien à voir entre eux.

L'échelle des snaps est obtenue à l'aide des estimateurs de temps musicaux et du premier temps des mesures de Davies *et al.* [Dav07, SDP09]. Cette échelle est synchrone au premier temps de chaque mesure et telle que la période des snaps est proche de 1 s.

L'estimation des frontières est réalisée par la méthode d'optimisation de coût mise en oeuvre par l'algorithme de Viterbi décrit dans la partie 4.1.5.3, de la même manière qu'en 2010. Le coût de segmentation est défini par les équations 4.1 et 4.2 où Φ est une fonction de similarité choisie de manière à associer un faible coût aux segments dont la séquence de descripteurs se répète ailleurs dans le morceau. Soit $X = \{x_1, x_2, ..., x_N\}$ la séquence de descripteurs d'un morceau. Soit $\{x_{t_k}, ..., x_{t_k+m_k}\}$ la séquence de descripteurs associée au segment s_k pour $1 < k < K$, on pose

$$\Phi(s_k) = \min_{\theta \in Z_k} \left\{ \sum_{p=0}^{m_k-1} 1 - \delta(x_{t_k+p}, x_{\theta+p}) \right\} \tag{5.10}$$

où δ correspond au symbole de Kronecker : $\delta(x_i, x_j) = 1$ si $x_i = x_j$ et $\delta(x_i, x_j) = 0$ sinon. Nous considérons que $Z_k = [1, t_k - m_k] \cup [t_k + m_k, N]$ afin d'éviter les comparaisons d'ordre interne au segment.

10. http ://versamus.inria.fr/

La contrainte de régularité Ψ a été choisie parmi trois éléments représentatifs de la famille de fonctions Ψ_α décrite dans la partie 4.1.3 : $\{\Psi_\alpha\}_{\alpha=0.5,1,2}$. Notre expérience d'annotation nous permet de fixer empiriquement τ à 16 snaps. Les paramètres $\alpha = 0.5$ et $\lambda = 0.13$ ont été réglés sur la base MIREX10 (IRISA).

L'algorithme d'estimation des étiquettes structurelles consiste en une version plus ancienne de l'approche par automates à états finis décrite dans la partie 4.2.1.3 [11]. On utilise la distance d'édition [12] afin de comparer deux séquences de symboles associées deux segments. À chaque étape du groupement hiérarchique, on fusionne les branches d'automate les plus proches au sens de la distance de type saut minimal : il s'agit de considérer la distance la plus petite entre une séquence de symboles de la première branche et une séquence de symboles de la seconde.

5.3.2 Participants

CL1 utilise une approche multi-strate et mono-critère. Un ensemble d'estimations de la structure est obtenu à partir d'un groupement hiérarchique à deux niveaux des deux types de descripteurs considérés (MFCCs et vecteurs de chroma). Les estimations sont combinées sous la forme d'une matrice dont les coefficients reflètent les correspondances entre les étiquettes des estimations au cours du temps. Les estimations des frontières structurelles sont issues de la segmentation de la matrice à l'aide d'un critère d'homogénéité basée sur une décomposition par NMF [CL11b].

GP6 correspond en majeure partie à l'algorithme GP7 des campagnes d'évaluations de 2010. MHRAF3 correspond à l'algorithme MHRAF2 soumis en 2010. Deux autres versions de ce même algorithme ont été soumises, et correspondent à MHRAF1 et MHRAF2. Les détails sur les algorithmes GP6, MHRAF1 et MHRAF2 n'ont cependant pas été publiés.

5.3.3 Résultats obtenus

Résultats concernant l'estimation des frontières structurelles Le tableau 5.5 permet d'observer, de la même manière qu'en 2010, que les performances pour l'estimation des frontières structurelles sont globalement plus élevées pour MIREX10 (IRISA) en comparaison de MIREX09. Celles obtenues pour l'algorithme IRISA11 permettent de mettre en valeur l'utilisation du critère de répétition et d'une fonction de régularité non-convexe ($\alpha = 0.5$), en particulier pour la tolérance de 0.5 s. Ces performances nous permettent d'atteindre la tête du classement selon les F_{br}, mais restent néanmoins proches de celles obtenues lors de la campagne de MIREX 2010.

IRISA11 a été évalué sur les bases Eurovision (Test) et Quaero Test09+10 et Test11. Le tableau 5.6 permet d'observer des résultats légèrement meilleurs en terme de F-mesure moyenne sur la base Quaero Test09+10 (IRISA), annotée selon la méthodologie du chapitre 3, contrairement à la base Quaero Test09+10 annotée par l'IRCAM. Les résultats obtenus sur l'Eurovision sont globalement meilleurs que ceux obtenus sur

11. Nous nous sommes récemment aperçus que la manière dont était calculée la probabilité des automates omettait certaines transitions entre états. Cette erreur a été propagée dans le système IRISA12 mais a été corrigée dans le cadre du diagnostic de sa version améliorée dans la partie 6.4.

12. Nous utilisons actuellement l'implémentation de Miguel Castro, disponible sur le site `http://www.mathworks.com/matlabcentral/fileexchange/213-editdist-m`. On associe aux opérations d'insertion, de suppression et de substitution de symboles le même poids, égal à 1.

Base MIREX09									
	segmentation (tol = 0.5s)			segmentation (tol = 3s)			étiquetage		
Participants	F_{br}	P_{br}	R_{br}	F_{br}	P_{br}	R_{br}	pF	pP	pR
CL	15.10	15.58	15.68	40.98	41.78	42.74	53.61	49.12	65.06
GP6	17.44	13.66	25.38	48.67	38.19	70.41	49.84	67.66	42.50
MHRAF1	20.63	20.88	21.70	51.90	52.46	54.75	54.45	47.43	71.97
MHRAF2	20.62	20.63	22.12	52.32	51.96	56.43	54.74	48.87	69.22
MHRAF3	18.53	20.17	18.07	50.79	54.86	49.88	**55.46**	50.51	67.86
IRISA11	**23.07**	**20.15**	**28.63**	**53.34**	**46.59**	**66.38**	**48.66**	**58.73**	**46.49**
Base MIREX10 (IRISA)									
	segmentation (tol = 0.5s)			segmentation (tol = 3s)			étiquetage		
Participants	F_{br}	P_{br}	R_{br}	F_{br}	P_{br}	R_{br}	pF	pP	pR
CL	23.07	30.84	19.06	43.36	57.05	36.28	-	-	-
GP6	18.78	17.49	21.02	53.37	49.80	59.53	-	-	-
MHRAF1	27.72	38.34	22.48	52.82	72.23	43.00	-	-	-
MHRAF2	26.86	35.29	22.39	56.26	73.82	46.87	-	-	-
MHRAF3	20.30	31.63	15.31	48.64	75.19	36.90	-	-	-
IRISA11	**32.38**	**32.64**	**33.23**	**61.20**	**62.20**	**62.33**	-	-	-

TABLEAU 5.5 – Résultats de l'évaluation de l'estimation des frontières structurelles et de la structure complète sur les bases MIREX09 et MIREX10 (IRISA), obtenus dans le cadre de la campagne MIREX 2011. Les *boundary hit rates* F_{br}, P_{br} et R_{br} ont été calculées pour les tolérances tol=0.5 s et 3 s. Les valeurs du système soumis apparaissent en gras. Les valeurs en bleu correspondent aux valeurs maximales sur l'ensemble des participants.

Quaero, ce qui peut s'expliquer par le fait que les morceaux de la première base sont relativement plus homogènes dans leur structure en comparaison de la seconde.

Résultats concernant l'estimation de la structure complète Les mesures de pF, pP et pR sont regroupées dans le tableau 5.5 pour les différents participants. Ceci permet d'observer que ces performances sont de l'ordre de celles obtenues en 2010. Les résultats de IRISA11 qui utilise cette fois un critère de répétition pour l'étiquetage, peuvent être expliqués par le fait que l'approche n'est pas robuste aux transpositions (contrairement à MHRAF). De plus, le processus de factorisation des branches d'automate est peu robuste à l'imprécision des estimations des frontières et aux distorsions temporelles, comme les ajouts et les troncatures que peuvent subir les blocs structurels recherchés (MHRAF et GP6 utilisent des méthodes d'alignement de séquences de vecteurs de descripteurs). Malgré les performances de CL1 concernant les frontières structurelles, cet algorithme parvient à obtenir une mesure de pF comparable celles des autres algorithmes ce qui souligne l'efficacité de son approche par homogénéité pour l'étape d'étiquetage. Notons qu'il combine des descripteurs de type timbral et tonal (MFCCs et chroma), ce qui peut contribuer à ses performances.

La campagne Quaero de 2011 a fait intervenir les algorithmes GP7 (version de GP6 de 2010) et IRISA11 (*cf.* tableau 5.6). Le classement entre ces deux algorithmes sur les bases Quaero Test09+10 et Test11 est le même que pour IRISA11 et GP6 à MIREX.

Base Eurovision (Test)				
Participants	F	P	R	score de modélisation
IRISA11	**53.28**	**57.15**	**51.59**	-
Base Quaero Test09+10 (IRISA)				
Participants	F	P	R	score de modélisation
IRISA11	**48.53**	**51.48**	**47.85**	-
GP7	41	41	42	-
Base Quaero Test11 (IRISA)				
Participants	F	P	R	score de modélisation
IRISA11	**45.61**	**46.24**	**49.38**	-
GP7	49	50	52	-
Base Quaero Test09+10 (IRCAM)				
Participants	F	P	R	score de modélisation
IRISA11	**43.77**	**49.85**	**41.80**	**51.74**
GP7	53	54	56	62
Base Quaero Test11 (IRCAM)				
Participants	F	P	R	score de modélisation
IRISA11	**43.44**	**45.32**	**45.09**	**49.59**
GP7	56	60	57	58

TABLEAU 5.6 – Résultats de l'évaluation de l'estimation des frontières structurelles et de la structure complète sur les différentes bases de test de la campagne Quaero 2011. F, P et R sont calculés pour une tolérance de 2 s.

5.4 Campagnes d'évaluation MIREX et Quaero de 2012

5.4.1 Contexte méthodologique et algorithme proposé

Les algorithmes d'estimation de structure de l'état de l'art fondent généralement l'estimation des frontières structurelles sur des considérations externes aux segments eux-mêmes : un segment est homogène, détecté par la rupture de l'homogénéité par rapport à ses voisins. un segment est répété, ou un segment est répété au cours du morceau. La conception de l'algorithme que nous avons soumis aux campagnes d'évaluation de 2012 est motivée par la volonté d'utiliser un modèle d'organisation interne aux segments structurels pour l'estimation de leurs frontières. Il s'agit du modèle système/contraste présenté dans la partie 3.5.3.

L'algorithme IRISA12 soumis cette année aux évaluations MIREX et Quaero reprend l'approche utilisée en 2011 en remplaçant le critère de répétition par le critère morphologique défini dans la partie 4.1.2.4.

Nous utilisons cette fois des descripteurs numériques (vecteurs de chroma) pour décrire les morceaux de musique dans le cadre de l'estimation des frontières structurelles. Ces descripteurs sont les *Chroma Pitch* (CP) générés par la Chroma Toolbox de Muller et Ewert [ME11]. Ceux-ci sont exprimés à l'échelle des snaps en associant à chacun d'entre eux le vecteur de chroma moyen issu des vecteurs contenus dans une fenêtre dont la durée égale la période des snaps et qui est centrée sur le snap courant. Les snaps sont estimés de la même manière qu'en 2011.

Nous utilisons le même modèle de contrainte de régularité qu'en 2011, avec une pulsation structurelle de $\tau = 16$ snaps. Le coût de segmentation est défini par les équations 4.1 et 4.2 où Φ est le coût issu du critère morphologique décrit par l'équation 4.10 et avec $\lambda_1 = 1$, $\lambda_2 = 0.04$, $\lambda = 0.41$ et $\alpha = 0.93$. Ces paramètres sont réglés sur notre base d'étude MIREX10 (IRISA).

Les segments obtenus sont étiquetés à l'aide de l'algorithme de regroupement hiérarchique utilisé en 2011 pour lequel on a effectué plusieurs modifications. Le calcul des probabilités des automates décrit dans la partie 4.2 est toujours effectué à partir d'une description symbolique du morceau de musique, mais l'ordre de fusion des branches d'automate est établi à l'aide des vecteurs de chroma dans l'optique de rendre l'étiquetage robuste aux changements de tonalité. On compare ainsi deux segments en calculant la distance entre la séquence de vecteurs de chroma du premier avec la séquence de chroma du second pour toutes les transpositions possibles. Transposer une séquence de vecteurs de chroma d'un demi-ton correspond à effectuer une permutation circulaire des coefficients des de chacun de ces vecteurs. La mesure de distance utilisée est la *stripe distance* [PK08a] et correspond à une mesure d'alignement temporel (*cf.* DTW partie 2.6). La distance entre deux segments correspond à la *stripe distance* minimale sur l'ensemble des 12 transpositions possibles. De la même manière que pour l'algorithme de 2011, la distance entre deux groupes de segments (c'est-à-dire deux branches d'automates) correspond à la plus petite distance entre un élément du premier groupe et un élément du second.

On choisit de plus de considérer chaque segment par les trois-quarts de leur séquence de descripteurs dans le cas où leur taille est d'au moins 16 snaps. Ceci permet d'éviter de considérer la partie des blocs structurels la plus susceptible de varier au cours du morceau (le contraste), en considérant qu'ils soient constitués de quatre éléments morphologiques.

La séquence de symboles est obtenue par quantification des vecteurs de chroma par un algorithme de Quantification Vectorielle, avec un nombre de classes de vecteurs de chroma empiriquement fixé à 16. On utilise la méthode LBG ou Lloyd généralisé, initialisée par division récursive des données selon le barycentre ou *splitting* [Gra84].

L'algorithme IRISA12_2 soumis à la campagne Quaero correspond à IRISA12 avec un paramétrage différent ($\lambda_1 = 1$, $\lambda_2 = 0.15$, $\lambda = 0.49$ et $\alpha = 1.11$) réglé sur l'ensemble de développement de la campagne Quaero 2012 (regroupant les bases MIREX10 (IRISA, Dev) et Eurovision (Dev)).

5.4.2 Participants

Les soumissions KSP1, KSP2, KSP3, SP1, IRCAM1, IRCAM2 et IRCAM3 correspondent à plusieurs versions du même algorithme. L'estimation des frontières structurelles est effectuée via un critère d'homogénéité sur des descripteurs de timbre (les MFCCs et leurs moments) et de type tonal (les *multi-probe histogram* dérivés des vecteurs de chroma) combinés en une matrice de similarité. Ce critère d'homogénéité correspond à la fonction de nouveauté décrite dans la partie 2.6 à laquelle on applique un seuil adaptatif. Les segments obtenus sont étiquetés à l'aide d'un critère de répétition. Chaque segment est caractérisé par les coefficients de la matrice de similarité décomposée par un algorithme NMF. SP1, KSP1 (=IRCAM2), KSP3 (=IRCAM1) et IRCAM3 diffèrent par l'ordre de la décomposition par NMF (respectivement 4, 6, 8 ,10). KSP2 correspond à une version de l'algorithme SP1 n'utilisant pas les descripteurs *multi-probe histogram* [KSG12].

SMGA1 et SMGA 2 sont deux versions du même algorithme qui fonde l'estimation des frontières structurelles sur un critère de répétition conceptuellement proche du critère de détection des ruptures de répétition décrit dans la partie 4.1.2.2. Les descripteurs utilisés dérivent des vecteurs de chroma : chaque descripteur correspond au vecteur de chroma de son instant correspondant auquel on concatène un ensemble d'échantillons de la séquence de chromas qui le précède. Ceci permet de modéliser la mémoire à court-terme d'un auditeur. L'estimation des frontières structurelles consiste ensuite à calculer une matrice de similarité à partir de ces descripteurs et à localiser les instants séparant les séquences de coefficients sous-diagonaux de forte similarité. Les détails concernant l'estimation des étiquettes structurelles n'ont pas encore été publiés [SMPA12].

MHRAF2 correspond au même algorithme que ceux nommés MHRAF1, MHRAF2 et MHRAF3 lors des campagnes de 2010 et 2011 avec un réglage différent de ses paramètres.

Les détails du système OYZS1 n'ont pas encore été publiés.

5.4.3 Résultats obtenus

Résultats concernant l'estimation des frontières structurelles Le tableau 5.7 permet d'observer une augmentation des performances globales obtenues sur MIREX09 et MIREX10 (IRISA) en comparaison des années précédentes pour l'estimation des frontières. Celles-ci mettent en avant les algorithmes SMGA et KSP pour la tolérance de 3 s, et KSP pour celle de 0.5 s. Les approches de ces algorithmes innovent principalement par leurs descripteurs : pour KSP, les *multi-probe histograms* représentant l'évolution du contenu tonal plutôt que le contenu tonal lui-même et les descripteurs

de SMGA prennent en compte une modélisation de la mémoire à court-terme. On observe de nouveau que les performances sur MIREX10 (IRISA) sont plus élevées que celles obtenues sur MIREX09. L'algorithme IRISA12 obtient des résultats comparables à ceux des algorithmes IRISA10_1&2 et IRISA11, présentés les années précédentes, sur ces deux bases.

Deux nouvelles bases ont été utilisées cette année dans MIREX : les annotations originales de MIREX10 produites avec RWC Pop par l'AIST et MIREX12. Les résultats obtenus par les différents algorithmes sont regroupés dans le tableau 5.8. L'observation des résultats pour les bases MIREX10 (IRISA et AIST), regroupés dans la figure 5.3, permet de constater que les mesures de F_{br} moyennes sur MIREX10 (AIST) sont légèrement plus faibles que pour MIREX10 (IRISA) et pour presque l'ensemble des algorithmes (seuls OYZS1 et IRISA12 voient leur F_{br} moyenne à 3 s augmenter de quelques pourcent). Ceci peut s'interpréter en terme de cohérence des annotations de MIREX10 (IRISA) par rapport à celles de MIREX10 (AIST).

Les tendances des résultats obtenus sur la nouvelle base MIREX12 sont différentes : les F_{br} moyennes à 0.5 s baissent mais restent comparables à celles des autres bases. En revanche, l'ensemble des algorithmes voient leur F_{br} moyennes à 3 s passer en dessous de 50%. Le déséquilibre des mesures P_{br} et R_{br} moyennes (de l'ordre de 30% avec $P_{br} < R_{br}$) traduit l'estimation d'un nombre trop élevé de frontières structurelles dans le cas des algorithmes KSP et SMGA. IRISA12 possède aussi ce comportement sur cette base, tandis que OYZS1 tend à estimer un nombre de frontières structurelles plus faibles que celles de référence et que MHRAF1 voit ses mesures P_{br} et R_{br} moyennes équilibrées. Ceci peut s'interpréter par le fait que les annotations de référence correspondent en général à une échelle structurelle plus grossière que les annotations estimées, comme semblent le montrer les aperçus des résultats disponibles sur le site de MIREX [13]. Les aperçus de trois morceaux de MIREX12 sont représentés dans la figure 5.2 (les références n'ont pas été rendues publiques).

Les algorithmes IRISA12_2 et IRCAM1/2/3, correspondent aux algorithmes IRISA12 et KSP1 pour lesquels les paramètres sont réglés différemment. Le tableau 5.9 permet d'observer que leur classement correspond à celui observé dans le cadre de MIREX. Les résultats obtenus par IRISA12_2 sur la base Eurovision (Test) sont globalement comparables à ceux du système IRISA11. Le F_{br} diminue d'un pourcent, et l'écart entre P_{br} et R_{br} se creuse avec $P_{br} > R_{br}$, ce qui implique que le système IRISA12_2 a tendance à estimer un nombre de frontières plus faible que le nombre de frontières de référence. Les algorithmes IRCAM n'ont pas été évalués sur la base Eurovision.

Résultats concernant l'estimation de la structure complète Les performances moyennes des différents algorithmes sur MIREX09, MIREX10 (AIST) et MIREX12 sont répertoriées dans le tableau 5.8. L'algorithme d'étiquetage de IRISA12 a été soumis à MIREX à titre exploratoire. Les modifications que l'on a effectué par rapport à l'algorithme IRISA11 de 2011 visent à évaluer le concept de S&C plutôt qu'à obtenir les meilleures performances possibles. On observe que les performances obtenues par IRISA12 sont légèrement plus élevées que celles obtenues par IRISA11 sur cette même base (il faut tenir compte de l'amélioration de l'estimation des frontières structurelles). Ceci peut venir du fait que la comparaison entre les segments structurels est basée sur leurs trois premiers quarts, et est plus robuste changements de tonalité. Une

13. `http://nema.lis.illinois.edu/nema_out/mirex2012/results/struct/sal/comparisonplots.html#segmentssalami000000`

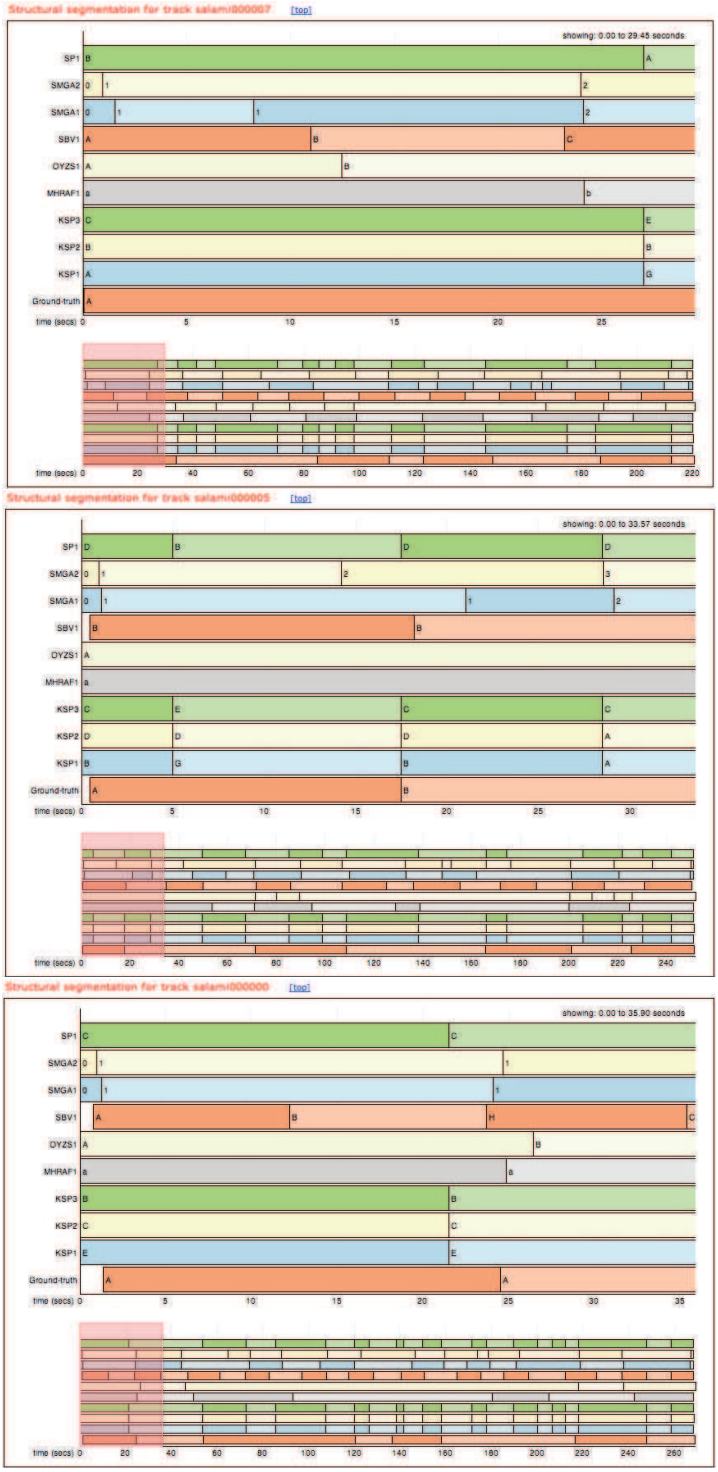

FIGURE 5.2 – Aperçu des estimations de structure des algorithmes soumis à MIREX 2012 pour trois morceaux de musique. Dans chaque cas, la structure la plus basse (orange) correspond à l'annotation de référence (les références n'ont pas été rendues public). Ces figures sont issues du site de MIREX 2012.

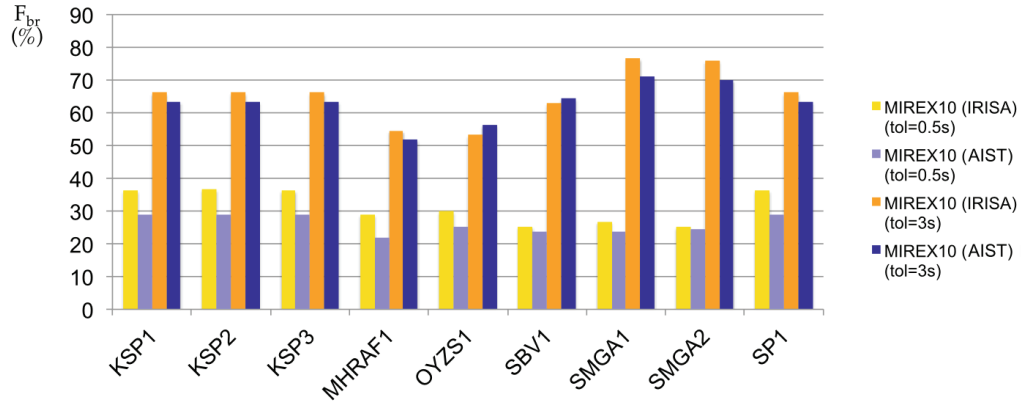

FIGURE 5.3 – Comparaison des mesures de F_{br} pour les tolérances de 0.5 s et 3 s obtenues avec les algorithmes de MIREX 2012 sur les bases MIREX10 (IRISA et AIST).

étude supplémentaire est nécessaire afin de diagnostiquer ce résultat. Les algorithmes MHRAF, KSP et SMGA sont tous les trois basés sur des critères de répétition et des descripteurs de type tonal (chromas), sont les plus performants sur les trois bases. Leurs résultats sont les meilleurs obtenus à l'échelle des trois années pour la structure complète mais aussi pour les frontières. Au vu des performances obtenues à l'échelle des campagnes de 2010 à 2012 (WB1 et MHRAF2 en 2010 et CL1 en 2011), il est difficile de privilégier un critère de répétition devant un critère d'homogénéité afin d'améliorer significativement l'estimation des étiquettes structurelles.

Les scores de modélisation moyens obtenus dans le cadre de Quaero sont répertoriés dans le tableau 5.9 et affichent les mêmes tendances qu'à MIREX.

5.5 Observations générales

Les algorithmes que nous avons soumis donnent des résultats comparables à l'état de l'art, en particulier pour l'estimation des frontières structurelles sur les bases MIREX09, MIREX10 (IRISA) et MIREX10 (AIST). Ceci permet de montrer la pertinence de l'utilisation d'une contrainte de régularité qui constitue leur composante commune. La méthodologie d'annotation de la base MIREX10 (IRISA) prend en compte une hypothèse de régularité, ce qui explique que les performances obtenues soient plus élevées sur cette base en comparaison des deux autres. Les annotations de la base MIREX12 semblent quant à elles beaucoup plus irrégulières.

Ces campagnes d'évaluation permettent de situer les algorithmes de l'état de l'art entre eux par rapport à un ensemble de bases donné. On peut noter que la marge de progression reste importante, notamment au niveau de l'estimation des frontières : les meilleures performances obtenues en terme de F-mesure sur l'ensemble des bases sont de l'ordre de 35% pour la tolérance de 0.5 s et 75% pour la tolérance de 3 s.

Il est cependant assez difficile d'évaluer plus précisément les avantages et les inconvénients des hypothèses et des composantes liées de chaque algorithme dans le cadre de ces campagnes. Ceci est lié à la variété des descripteurs utilisés ainsi qu'à la consistance des annotations de référence. En effet, les descripteurs sur lesquels les algorithmes fondent leur analyse sont calculés par des outils différents et sont exprimés à différentes échelles (trame de durée fixe, temps et/ou mesures musicales...). Ensuite, la variété

Base MIREX09									
	segmentation (tol = 0.5s)			segmentation (tol = 3s)			étiquetage		
Participants	F_{br}	P_{br}	R_{br}	F_{br}	P_{br}	R_{br}	pF	pP	pR
KSP1	**28.16**	24.14	35.77	59.13	50.64	75.00	55.40	63.26	52.96
KSP2	27.99	23.98	35.57	59.07	50.55	74.98	54.46	47.33	69.78
KSP3	**28.16**	24.14	35.77	59.13	50.64	75.00	57.18	56.61	62.26
MHRAF1	22.03	22.50	22.83	52.51	53.42	54.65	55.64	47.66	75.22
OYZS1	19.13	24.97	17.53	44.07	54.89	41.06	46.37	51.13	53.21
IRISA12	**22.67**	**21.61**	**25.02**	**55.42**	**52.70**	**61.19**	**51.47**	**53.62**	**55.55**
SMGA1	22.82	20.51	26.65	**64.49**	57.95	75.05	**65.28**	61.80	74.64
SMGA2	20.19	18.29	23.47	59.69	53.82	69.42	63.33	64.57	66.87
SP1	**28.16**	24.14	35.77	59.13	50.64	75.00	55.14	47.07	73.10
Base MIREX10 (IRISA)									
	segmentation (tol = 0.5s)			segmentation (tol = 3s)			étiquetage		
Participants	F_{br}	P_{br}	R_{br}	F_{br}	P_{br}	R_{br}	pF	pP	pR
KSP1	36.25	37.61	36.21	66.08	68.68	65.83	-	-	-
KSP2	**36.61**	38.02	36.55	66.08	68.68	65.83	-	-	-
KSP3	36.25	37.61	36.21	66.08	68.68	65.83	-	-	-
MHRAF1	28.91	40.38	23.19	54.49	75.78	43.76	-	-	-
OYZS1	29.89	41.63	24.30	53.12	73.89	43.25	-	-	-
IRISA12	**25.31**	**28.35**	**23.31**	**62.81**	**69.75**	**58.36**	-	-	-
SMGA1	26.78	28.67	25.58	**76.57**	81.58	73.52	-	-	-
SMGA2	25.16	27.05	23.95	75.86	8130	72.47	-	-	-
SP1	36.25	37.61	36.21	66.08	68.68	65.83	-	-	-

TABLEAU 5.7 – Résultats de l'évaluation de l'estimation des frontières structurelles et de la structure complète sur les bases M09 et M10_I obtenus dans le cadre de la campagne MIREX 2012. Les *boundary hit rates* F_{br}, P_{br} et R_{br} ont été calculées pour les tolérances tol=0.5 s et 3 s. Les valeurs du système soumis apparaissent en gras. Les valeurs en bleu correspondent aux valeurs maximales sur l'ensemble des participants.

Base MIREX10 (AIST)									
	segmentation (tol = 0.5s)			segmentation (tol = 3s)			étiquetage		
Participants	F_{br}	P_{br}	R_{br}	F_{br}	P_{br}	R_{br}	pF	pP	pR
KSP1	**28.94**	29.60	29.42	63.40	64.37	64.83	60.26	65.35	57.69
KSP2	**28.94**	29.60	29.42	63.40	64.37	64.83	58.30	49.85	73.94
KSP3	**28.94**	29.60	29.42	63.40	64.37	64.83	60.45	57.71	66.21
MHRAF1	21.99	30.03	17.93	51.80	70.05	42.42	58.30	51.99	71.60
OYZS1	25.22	33.83	21.00	56.38	76.37	46.83	50.71	61.07	46.81
IRISA12	**23.67**	**25.80**	**22.64**	**64.37**	**70.02**	**61.57**	**53.50**	**61.77**	**50.47**
SMGA1	23.59	24.69	23.19	**71.01**	74.11	70.07	67.52	70.38	67.49
SMGA2	24.56	25.84	24.00	70.01	73.36	68.79	**68.83**	74.64	66.64
SP1	**28.94**	29.60	29.42	63.40	64.37	64.83	56.22	46.02	76.53
Base MIREX12									
	segmentation (tol = 0.5s)			segmentation (tol = 3s)			étiquetage		
Participants	F_{br}	P_{br}	R_{br}	F_{br}	P_{br}	R_{br}	pF	pP	pR
KSP1	27.87	22.34	43.68	49.02	39.21	76.71	50.19	66.53	44.64
KSP2	28.60	22.91	44.86	48.99	39.15	76.76	52.83	55.03	57.92
KSP3	27.89	22.37	43.71	49.06	39.24	76.76	53.09	61.20	52.61
MHRAF1	18.79	19.44	19.92	42.29	44.46	44.02	57.22	56.40	67.23
OYZS1	**28.74**	45.80	25.27	43.68	64.09	39.70	50.06	58.17	59.54
IRISA12	**15.66**	**13.59**	**20.95**	**43.44**	**37.81**	**57.44**	**45.96**	**62.71**	**42.50**
SMGA1	19.24	15.63	28.16	**49.20**	40.40	70.28	**58.09**	67.62	58.26
SMGA2	17.82	14.60	25.72	47.89	39.59	67.79	52.82	72.85	47.12
SP1	27.89	22.37	43.71	49.06	39.24	76.76	55.43	54.90	63.95

TABLEAU 5.8 – Résultats de l'évaluation de l'estimation des frontières structurelles et de la structure complète sur les bases MIREX10 (AIST) et MIREX12 obtenus dans le cadre de la campagne MIREX 2012. Les *boundary hit rates* F_{br}, P_{br} et R_{br} ont été calculées pour les tolérances tol=0.5 s et 3 s. Les valeurs du système soumis apparaissent en gras. Les valeurs en bleu correspondent aux valeurs maximales sur l'ensemble des participants.

Base Eurovision (Test)				
Participants	F	P	R	score de modélisation
IRISA12_2	**52.37**	**58.69**	**48.99**	-
Base Quaero Test09+10 (IRISA)				
Participants	F	P	R	score de modélisation
IRISA12_2	**46.42**	**50.49**	**44.96**	-
IRCAM1/2/3	56.0	59.5	55.5	-
Base Quaero Test 11 (IRISA)				
Participants	F	P	R	score de modélisation
IRISA12_2	**45.79**	**53.79**	**43.45**	-
IRCAM1/2/3	55	59	55	-
Quaero Test09+10 (IRCAM)				
Participants	F	P	R	score de modélisation
IRISA12_2	**43.01**	**45.71**	**43.96**	**54.3**
IRCAM1/2/3	55.0	55.0	59.5	60.2/60.9/61.1
Base Quaero Test 11 (IRCAM)				
Participants	F	P	R	score de modélisation
IRISA12_2	**44.43**	**51.33**	**42.53**	**49.4**
IRCAM1/2/3	57	59	59	59.0/58.6/57.7
Base Quaero Test 12 (IRCAM)				
Participants	F	P	R	score de modélisation
IRISA12_2	**45.60**	**49.11**	**45.06**	**53.6**
IRCAM1/2/3	59	59	62	60.6/59.9/58.7

TABLEAU 5.9 – Résultats de l'évaluation de l'estimation des frontières structurelles et de la structure complète sur les bases de test de la campagne Quaero 2012. F, P et R sont calculés pour une tolérance de 2 s.

des méthodologies d'annotation considérées implique que l'on évalue la capacité des algorithmes à produire des estimations qui ressemblent à ce que peut annoter un annotateur humain sans assurer de cohérence méthodologique particulière. Ehmann évalue dans [EBD+11] la concordance entre les annotations de chaque morceau de la base MIREX12 associé à deux annotations structurelles de référence. Il obtient $pF = 62.9\%$ pour l'échelle "fine" et $pF = 72.1\%$ pour l'échelle "grossière" qui restent assez loin de 100%. Ceci permet de souligner l'utilité d'une méthodologie d'annotation cohérente pour l'évaluation à grande échelle des algorithmes.

5.6 Bilan et extension des algorithmes soumis pour l'estimation des frontières

Les trois algorithmes que nous avons présenté à MIREX ont des performances globalement comparables en terme de F_{br} moyenne comme nous le montre la figure 5.4. Cependant ces valeurs ne nous permettent pas d'évaluer si les systèmes produisent des estimations complémentaires pour retrouver les frontières de référence. Nous avons donc comparé les frontières estimées par chacun de nos systèmes sur MIREX10 (IRISA) à celles estimées par chacun des autres systèmes.

Comparaison des frontières estimées par les différents systèmes Les F_{br} moyens issus de la comparaison entre les systèmes IRISA10_1 et IRISA10_2 sont 96.44% pour tol=0.5 s et 96.65% pour tol=0.5 s, ce qui valide le fait que leurs estimations soient quasi-identiques sur la base MIREX10 (*cf.* tableau 5.11). L'observation des F_{br} pour chaque morceau par les histogrammes de la figure 5.5 permet de s'apercevoir que les estimations ne coïncident pas pour six morceaux sur les 100. La visualisation directe des estimations a montré que pour trois d'entre eux, l'un des systèmes n'avait pas estimé de frontières, et que pour les trois autres, l'un des systèmes avait soit segmenté à une échelle plus fine que l'autre, soit avait légèrement déplacé quelques frontières par rapport à l'autre. Ainsi l'on pourra s'interroger à l'avenir sur la variation des résultats produits selon les ordinateurs utilisés, notamment du point de vue de l'extraction des descripteurs et de l'estimation de la pulsation structurelle.

Les estimations des systèmes IRISA10_1 et IRISA10_2 étant identiques sur la quasi-totalité de MIREX10 (IRISA), considérons maintenant les F_{br} moyens comparant celles des systèmes IRISA10_1, IRISA11 et IRISA12 entre elles. Les valeurs obtenues sont inférieures à 30% pour tol=0.5 s et à 60% pour tol=3 s, ce qui souligne la différence entre les estimations produites par les différents systèmes. Les valeurs moyennes des P_{br} du tableau 5.11 sont comparables aux valeurs moyennes des R_{br} (il s'agit des valeurs duales de P_{br} : elles peuvent être obtenues en reprenant les valeurs du tableau en échangeant le rôle estimation/référence) ce qui montre que le nombre de frontières estimées est globalement comparable d'un système à l'autre. Ceci nous incite à considérer leur combinaison à l'avenir.

Étude d'un système d'estimation des frontières basé sur les sorties des systèmes IRISA soumis à MIREX Nous proposons d'aborder la question de la combinaison de ces systèmes par le biais d'un système simple effectuant l'union des frontières estimées par les systèmes IRISA10_1, IRISA11 et IRISA12 sous la contrainte

de régularité Ψ_α utilisée dans IRISA11 et IRISA12. Ce système est choisi pour la facilité et la rapidité de sa mise en oeuvre.

Il consiste en la recherche de la segmentation de plus bas coût par l'algorithme de Viterbi de la partie 4.1.5.3. Le coût de segmentation est de nouveau obtenu par les équations 4.1 et 4.2. Le coût Φ est déduit d'un critère audio $\phi_\cup$ issu de l'union des frontières estimées par IRISA10_1, IRISA11 et IRISA12. Nous ne considérons pas IRISA10_2 pour éviter de sur-représenter notre système de 2010. Les frontières sont exprimées avec une résolution temporelle de 0.5 s pour des raisons de temps de calcul. Nous calculons un critère audio associant à chaque unité de temps t le nombre de frontières structurelles contenu dans la fenêtre de taille w centrée sur t. Φ est obtenu par l'équation

$$\Phi = \max(\phi_\cup) - \phi_\cup. \tag{5.11}$$

Le coût de régularité est Ψ_α, et la pulsation structurelle τ correspond au nombre d'unités de temps contenu dans 16 s.

Ce système est réglé et évalué sur la base MIREX10 (IRISA) pour les tolérances de 0.5 s et 3 s dans la perspective de le comparer aux résultats des trois algorithmes dont on utilise les estimations. On considère la grille de valeurs suivante pour régler les paramètres de ce système, que l'on nomme IRISA_Tous par la suite : $\alpha \in [0, 3]$ avec un pas de 0.1, $\lambda \in [0, 1]$ avec un pas de 0.01 et $w \in$[0 s,8 s] avec un pas de 0.2 s.

Un réglage optimal des paramètres selon le F_{br} moyen pour tol=0.5 s permet d'obtenir F_{br}=35.84% pour cette tolérance. Le meilleur système pour cette tolérance (IRISA11) avait obtenu en moyenne $F_{br} = 32.38\%$. Lorsque les paramètres sont réglés de manière à maximiser le F_{br} moyen pour tol=3 s, on obtient F_{br}=66.83% pour cette tolérance. Le meilleur système pour cette tolérance (IRISA12) avait obtenu en moyenne $F_{br} = 62.81\%$. Les résultats sont détaillés dans le tableau 5.12 et les courbes d'évolution du F_{br} moyen maximal pour les deux tolérances et pour les différents paramètres sont représentées dans les figures 5.8, 5.7 et 5.6.

Les résultats de cette étude permettent de mettre en lumière le potentiel lié à la combinaison de nos différents systèmes. S'agissant d'un algorithme réalisant une fusion tardive des critères audio considérés, il serait intéressant de considérer à l'avenir leur fusion au sein de l'algorithme de segmentation, par leur expression dans un même cadre à la manière des systèmes IRISA10_1&2.

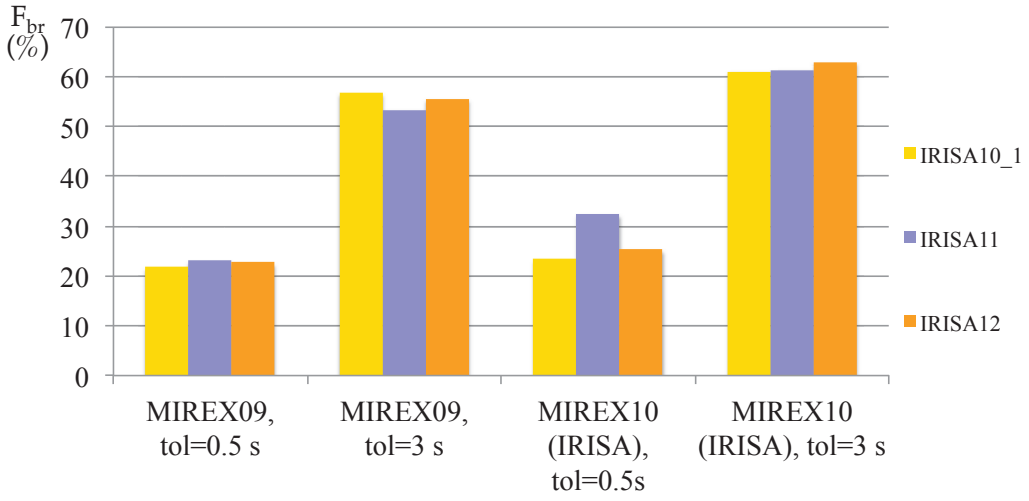

FIGURE 5.4 – Résumé des performances obtenues par les algorithmes IRISA10_1, IRISA11 et IRISA12 sur les bases MIREX09 et MIREX10 (IRISA).

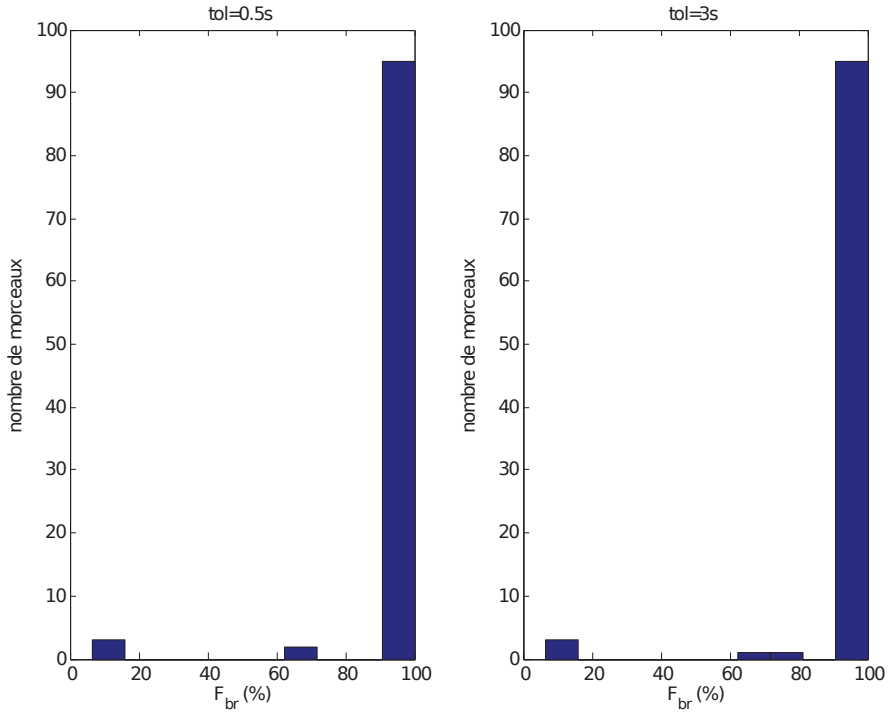

FIGURE 5.5 – Valeurs de F_{br} issues de la comparaison des estimations des systèmes IRISA10_1 et IRISA10_2 obtenues sur MIREX10 (IRISA).

F_{br} moyens (%), tol=0.5 s					
		estimation			
		IRISA10_1	IRISA10_2	IRISA11	IRISA12
référence	IRISA10_1	100.00	96.44	16.04	13.92
	IRISA10_2	96.44	100.00	15.67	13.85
	IRISA11	16.04	15.67	100.00	26.11
	IRISA12	13.92	13.85	26.11	100.00

F_{br} moyens (%), tol=3 s					
		estimation			
		IRISA10_1	IRISA10_2	IRISA11	IRISA12
référence	IRISA10_1	100.00	96.65	55.20	54.68
	IRISA10_2	96.65	100.00	54.61	54.36
	IRISA11	55.20	54.61	100.00	59.16
	IRISA12	54.68	54.36	59.16	100.00

TABLEAU 5.10 – F_{br} moyens comparant les frontières estimées sur la base MIREX10 (IRISA) par les différents systèmes que nous avons soumis aux campagnes d'évaluation MIREX, pour les tolérances tol=0.5 s (haut) et 3 s (bas).

P_{br} moyens (%), tol=0.5 s					
		estimation			
		IRISA10_1	IRISA10_2	IRISA11	IRISA12
référence	IRISA10_1	100.00	97.07	16.31	15.28
	IRISA10_2	97.03	100.00	16.00	15.23
	IRISA11	16.75	16.72	100.00	28.55
	IRISA12	13.54	13.84	24.17	100.00

P_{br} moyens (%), tol=3 s					
		estimation			
		IRISA10_1	IRISA10_2	IRISA11	IRISA12
référence	IRISA10_1	100.00	97.26	56.56	61.16
	IRISA10_2	97.27	100.00	56.14	60.94
	IRISA11	56.63	56.65	100.00	65.05
	IRISA12	51.51	51.90	54.51	100.00

TABLEAU 5.11 – P_{br} moyens comparant les frontières estimées sur la base MIREX10 (IRISA) par les différents systèmes que nous avons soumis aux campagnes d'évaluation MIREX, pour les tolérances tol=0.5 s (haut) et 3 s (bas). L'interversion des rôles d'estimation/référence permet d'obtenir les valeurs des R_{br} moyens sur la même base.

Paramètres réglés pour maximiser le F_{br} moyen pour tol=0.5 s									
Système	w^*	α^*	λ^*	$F_{br}(\%)$	$P_{br}(\%)$	$R_{br}(\%)$	$F_{br}(\%)$	$P_{br}(\%)$	$R_{br}(\%)$
IRISA_Tous	1	1.1	0.74	**35.84**	41.99	31.73	60.54	71.13	53.49
Paramètres réglés pour maximiser le F_{br} moyen pour tol=3 s									
Système	w^*	α^*	λ^*	$F_{br}(\%)$	$P_{br}(\%)$	$R_{br}(\%)$	$F_{br}(\%)$	$P_{br}(\%)$	$R_{br}(\%)$
IRISA_Tous	5.0	2.0	0.86	23.99	27.14	21.79	**66.83**	75.95	60.49

TABLEAU 5.12 – Performances moyennes optimales du système IRISA_Tous combinant les estimations des frontières de IRISA10_1, IRISA11 et IRISA12 sous la contrainte de régularité Ψ_α. Elles sont obtenues sur la base MIREX10 (IRISA) selon le F_{br} moyen pour tol=0.5 s (haut) et tol=3 s (bas).

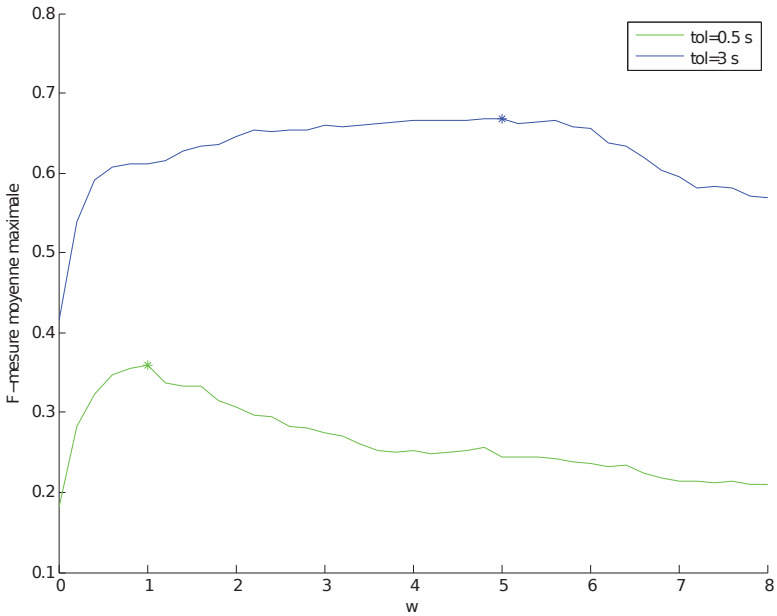

FIGURE 5.6 – Courbe d'évolution du maximum du F_{br} moyen obtenu par le système IRISA_Tous sur MIREX10 (IRISA) en fonction de la taille de fenêtre d'analyse w utilisée pour le calcul du critère audio issu de l'union frontières estimées par IRISA10_1, IRISA11 et IRISA12.

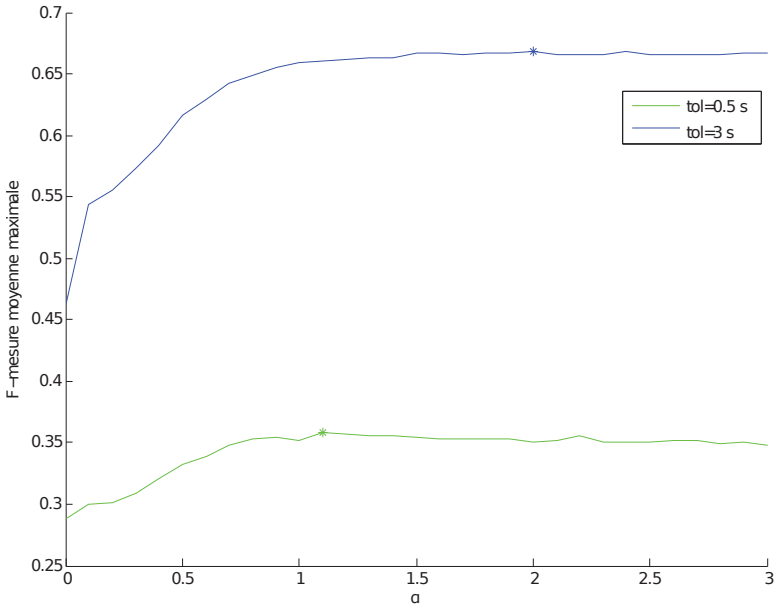

FIGURE 5.7 – Courbe d'évolution du maximum du F_{br} moyen obtenu par le système IRISA_Tous sur MIREX10 (IRISA) en fonction du paramètre de convexité α de la contrainte de régularité.

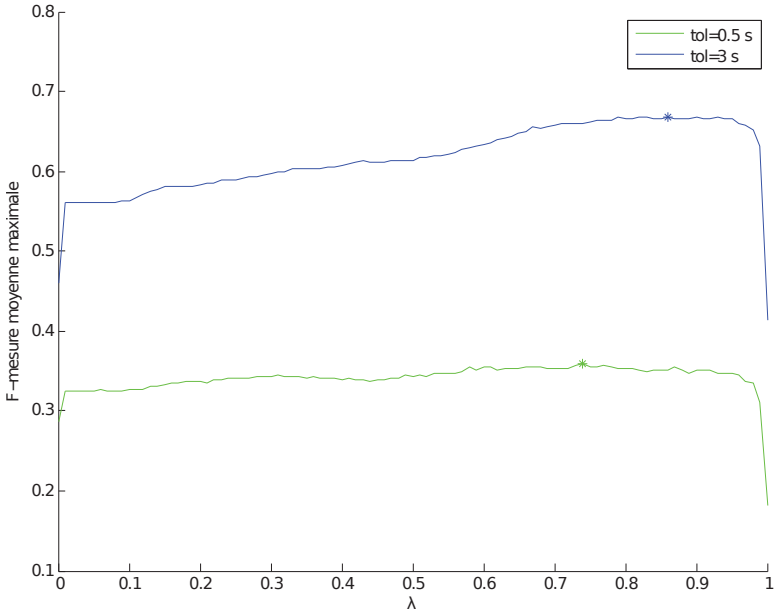

FIGURE 5.8 – Courbe d'évolution du maximum du F_{br} moyen obtenu par le système IRISA_Tous sur MIREX10 (IRISA) en fonction du paramètre de pondération λ de la contrainte de régularité.

5.7 Résumé du chapitre

Ce chapitre présente plusieurs systèmes d'estimation de structure basés sur les approches introduites dans le chapitre 4 et leurs performances obtenues dans le cadre des campagnes d'évaluation annuelles MIREX et Quaero. Ceci permet d'une part d'obtenir une première évaluation de ces approches et d'autre part de les comparer à l'état de l'art.

Les systèmes que nous avons soumis sont composés d'une phase d'estimation des frontières structurelles puis d'une phase d'estimation des étiquettes des segments obtenus.

En ce qui concerne la première phase, tous les systèmes réalisent l'optimisation du coût de segmentation à l'aide d'un algorithme de Viterbi, comme proposé dans le chapitre 4 et utilisent une contrainte de régularité structurelle. Le système de 2010 implémente une approche multicritère en combinant une version filtrée du critère de rupture d'homogénéité timbral, du critère de rupture de répétition du contenu tonal et du critère de détection d'événements du point de vue du timbre par une somme pondérée. Il utilise une première fonction de régularité asymétrique, convexe et minimale en la pulsation structurelle, estimée à partir du critère d'homogénéité qui s'avère être le plus régulier des trois. Le système de 2011 combine un critère de répétition calculé sur une représentation symbolique du contenu tonal et une contrainte de régularité issue de la famille de fonctions présentée dans le chapitre précédent et pour une pulsation structurelle fixée empiriquement à 16 snaps. Le système de 2012 combine un critère morphologique calculé sur les vecteurs de chroma et une contrainte de régularité issue de la même famille qu'en 2011.

Ces trois systèmes ont obtenu des performances comparables à celles des autres algorithmes de l'état de l'art pour l'estimation des frontières structurelles. Ceci permet de valoriser l'approche multicritère, l'utilisation d'un critère morphologique et d'une contrainte de régularité. Nous explorons leur combinaison par l'évaluation d'un système supplémentaire qui effectue l'union des estimations obtenues par les systèmes sous une contrainte de régularité. Ceci nous permet de mettre en évidence une amélioration significative de l'estimation des frontières sur une base issue de la méthodologie d'annotation du chapitre 3.

Une partie des bases que nous avons produites avec notre méthodologie d'annotation a été utilisée dans ces évaluations. On note dans le cas de MIREX que les performances obtenues par les différents systèmes soumis sont plus élevées sur cette partie. Ceci peut s'expliquer par le type de musique considéré et très probablement par le souci de cohérence de la méthodologie sur l'ensemble des annotations produites.

Concernant la seconde phase, les trois systèmes utilisent un regroupement hiérarchique des segments structurels estimés. Le système de 2010 fonde la comparaison de deux segments sur leur homogénéité timbrale, tandis que les systèmes de 2011 et 2012 utilisent la répétition de leur contenu tonal. Les performances obtenues sont en général plus faibles que celles de l'état de l'art, ce qui peut être lié à des problème de robustesse aux erreurs de localisation des frontières structurelles. L'étude des performances des systèmes de l'état de l'art ne permet pas de privilégier un critère particulier (répétition ou homogénéité) pour l'étiquetage automatique des segments structurels.

Chapitre 6

Diagnostic des approches pour l'estimation de la structure sémiotique

Ce chapitre présente un diagnostique des principaux modules qui mettent en oeuvre nos approches pour l'estimation de la structure sémiotique et que nous avons introduits dans le chapitre 4. Ceci vient compléter l'analyse des systèmes que nous avons soumis lors des campagnes MIREX et Quaero et dont nous avons présenté les performances dans le chapitre 5.

Nous étudions dans un premier temps le potentiel de plusieurs critères audio, considérés séparément puis conjointement, pour la segmentation. Le critère morphologique introduit dans la partie 4.1.2.4, soumis aux campagnes d'évaluation de 2012 à titre exploratoire et dont la formulation diffère de celle des autres critères étudiés, ne sera pas considéré dans ce chapitre. Les premières expériences font intervenir plusieurs paramètres réglés en utilisant les annotations de référence (cas oracle). Celles-ci nous permettent de disposer de premières performances auxquelles on pourra se référer dans la suite du chapitre. Nous évaluons ensuite l'impact de l'utilisation d'une contrainte structurelle dans le processus d'estimation des frontières par l'évaluation d'un système combinant le meilleur critère parmi ceux étudiés, avec la famille de contraintes de régularité introduite dans la partie 4.1.3. Enfin, nous étudions l'intérêt de considérer conjointement plusieurs critères audio et une contrainte de régularité en évaluant un système composé de ces caractéristiques.

Dans un second temps nous étudions le système d'étiquetage expérimental présenté dans la partie 4.2 en supposant les frontières structurelles connues. Nous considérons plusieurs versions de ce système afin d'étudier l'efficacité d'un critère auto-adaptatif par rapport à un critère de l'état de l'art et de mettre en lumière l'intérêt du modèle système-contraste introduit dans la partie 3.5.3.

6.1 Segmentation structurelle par analyse multicritère

Nous étudions ici le potentiel des critères introduits dans la partie 4.1.2.2 pour la segmentation, séparément puis conjointement, dans le cadre d'expérimentations oracles. Les systèmes étudiés sont réglés et testés sur le même ensemble de morceaux de musique.

6.1.1 Contexte expérimental

6.1.1.1 Corpus de morceaux

Nous utilisons la base RWC Pop décrite dans la partie 3.7 pour l'ensemble des expérimentations conduites dans ce chapitre. Ces morceaux sont originalement en stéréo et échantillonnés à 44100Hz. On utilise leur version mono, en calculant la moyenne des deux canaux. Nous nous référons aux frontières structurelles des annotations de référence que l'on a produites avec la méthodologie du chapitre 3 par l'expression "frontières de référence". Ces annotations correspondent à une version plus récente de la base MIREX10 (IRISA) ayant fait l'objet de révisions mineures suite à l'affinement de notre méthodologie d'annotation de la structure.

6.1.1.2 Métriques d'évaluation

L'évaluation de l'estimation des frontières structurelles est effectuée par le calcul des *boundary hit rates* F_{br}, P_{br} et R_{br}. On utilise le score de modélisation pour l'évaluation de l'estimation de la structure complète. Ces métriques sont définies dans la partie 5.1.2.

6.1.1.3 Descripteurs utilisés

Nous utilisons les descripteurs acoustiques MFCC et les vecteurs de chroma présentés dans la partie 4.1.5.1.

6.1.2 Étude des critères séparés pour la segmentation

Cette partie présente une étude de l'apport des critères audio introduits dans le chapitre 4.1.2.2 dans le cadre de l'estimation des frontières structurelles. Ces critères sont évalués séparément.

6.1.2.1 Protocole d'évaluation oracle

Pour chaque morceau, les critères de rupture d'homogénéité ϕ_H et de répétition ϕ_R et le critère de détection d'événements ϕ_E sont calculés sur les séquences de descripteurs MFCC et de vecteurs de chroma et exprimés à l'échelle des temps musicaux. Les courbes obtenues sont ensuite filtrées à l'aide du critère de Seck introduit dans la partie 4.1.5.1. Pour chaque critère audio obtenu, on estime la position des frontières structurelles en détectant les pics qui dépassent un seuil particulier. Afin d'avoir une idée des performances maximales associés aux critères considérés, nous nous plaçons dans un contexte oracle où ce seuil est choisi *a posteriori* : il s'agit de celui qui maximise le F_{br} parmi l'ensemble des seuils testés (on considère toutes les valeurs non-nulles du critère filtré) pour chaque morceau de la base. On considère les tolérances de 0.5 s et 3 s.

6.1.2.2 Résultats

Le tableau 6.1 répertorie les performances moyennes obtenues pour les trois critères calculés sur les descripteurs MFCC et les vecteurs de chroma sur l'ensemble de la base de RWC Pop. On remarque que le critère de rupture d'homogénéité et le critère de détection d'événements obtiennent de meilleurs résultats avec les MFCCs. Le critère

Seuil de sélection des pics réglé pour tol=0.5 s			
	tol = 0.5 s		
	$F_{br}(\%)$	$P_{br}(\%)$	$R_{br}(\%)$
Critère ϕ_{H_m}	**36.69**	47.94	33.87
Critère ϕ_{H_c}	24.88	30.98	30.29
Critère ϕ_{R_m}	9.22	17.76	8.95
Critère ϕ_{R_c}	**13.89**	13.90	21.13
Critère ϕ_{E_m}	**12.78**	12.33	24.97
Critère ϕ_{E_c}	10.52	9.14	24.64

Seuil de sélection des pics réglé pour tol=3 s			
	tol = 3 s		
	$F_{br}(\%)$	$P_{br}(\%)$	$R_{br}(\%)$
Critère ϕ_{H_m}	**64.19**	71.81	60.70
Critère ϕ_{H_c}	52.65	53.71	55.65
Critère ϕ_{R_m}	33.19	45.04	28.69
Critère ϕ_{R_c}	**42.76**	41.75	48.40
Critère ϕ_{E_m}	**58.68**	61.27	60.61
Critère ϕ_{E_c}	48.09	45.70	57.97

TABLEAU 6.1 – Valeurs moyennes des performances oracles obtenues sur RWC Pop pour les critères de rupture d'homogénéité ϕ_H, de répétition ϕ_R et de détection d'événements ϕ_E calculés sur les descripteurs MFCC (critères d'indice m) et les vecteurs de chroma (critères d'indice c). Le seuil de sélection des pics des critères filtrés est réglé de manière à maximiser la F-mesure pour une tolérance de 0.5 s (partie haute du tableau) ou de 3 s (partie basse du tableau).

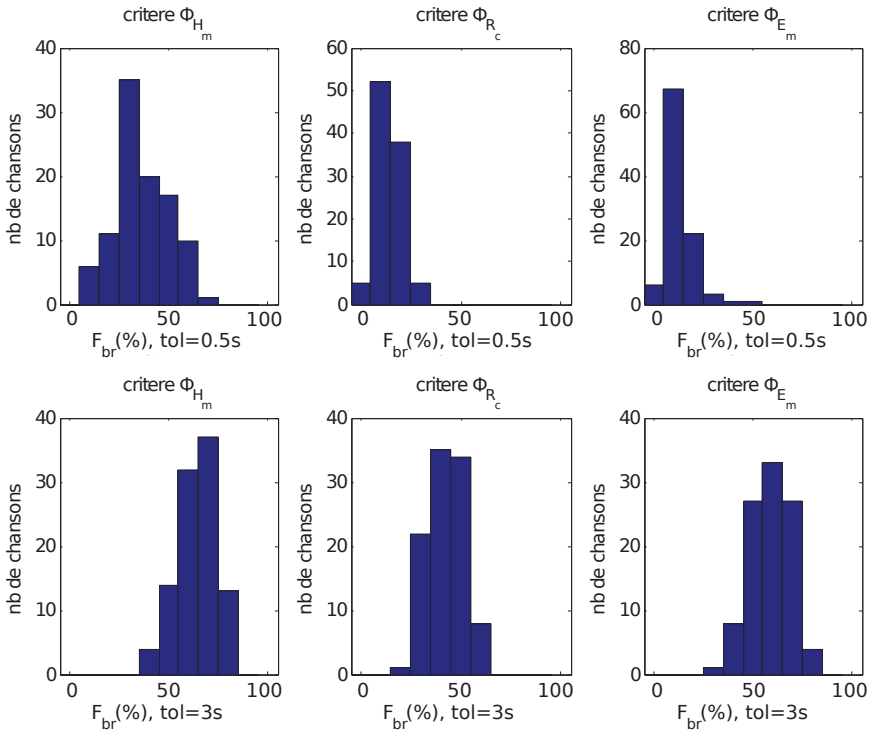

FIGURE 6.1 – Histogrammes des mesures F_{br} obtenues en oracle pour les 100 morceaux de RWC Pop et pour les trois meilleurs critères parmi ceux considérés. De gauche à droite, on a le critère de rupture d'homogénéité calculé sur les MFCCs, le critère de rupture de répétition calculé sur les vecteurs de chroma, puis le critère de détection d'événements calculé sur les MFCCs.

de rupture de répétitions quant à lui, donne de meilleurs résultats sur les vecteurs de chroma. Ces résultats nous incitent à sélectionner ces trois critères afin de l'étudier leur combinaison dans la suite de la thèse. Les résultats de ϕ_{H_m} sont largement supérieurs à ceux des autres critères. Ceci peut s'interpréter par le fait qu'une grande partie des morceaux de RWC Pop ont leurs frontières structurelles qui coïncident avec une rupture de timbre. Les autres critères obtiennent des performances plus faibles pour les deux tolérances considérées.

Les histogrammes de la figure 6.1 présentent la répartition des morceaux de la base étudiée en fonction de leur F_{br} oracle, et pour les trois critères les plus performants. On peut ainsi observer que cette répartition est concentrée autour du F_{br} oracle moyen, ce qui montre que les performances moyennes résument correctement le comportement mono-modal des performances à l'échelle des morceaux de musique dans sa globalité.

6.1.3 Étude des critères combinés pour la segmentation

Parmi les critères considérés, le critère d'homogénéité ϕ_H calculé sur les MFCC donne les meilleures performances moyennes en oracle sur la base RWC Pop. Cependant, toutes les frontières ne coïncident pas avec une rupture d'homogénéité de l'information contenue sur une strate musicale particulière. Par exemple, la figure 6.2 représente chaque critère calculé sur la séquence de descripteurs qui a permis d'obtenir les performances moyennes les plus hautes sur RWC Pop. Ces critères, bruts et filtrés,

Seuil de sélection des pics réglé pour tol=0.5 s			
	tol = 0.5 s		
	F_{br} (%)	P_{br} (%)	R_{br} (%)
$\phi_{H_m \cup R_c \cup E_m}$	17.11	10.57	56.33
Seuil de sélection des pics réglé pour tol=3 s			
	tol = 3 s		
	F_{br} (%)	P_{br} (%)	R_{br} (%)
$\phi_{H_m \cup R_c \cup E_m}$	37.32	24.81	78.52

TABLEAU 6.2 – Performances oracles moyennes sur RWC Pop, issues de l'union des frontières obtenues par sélection du seuil optimal pour ϕ_{H_m}, ϕ_{R_c} et ϕ_{E_m} pour les tolérances tol=0.5 s (partie haute du tableau) ou de 3 s (partie basse du tableau).

sont mis en regard de l'annotation de référence obtenue à l'aide de la méthodologie développée au chapitre 3. On observe que les trois critères apportent des informations complémentaires sur la position des frontières structurelles recherchées : par exemple, les trois frontières de référence à 80, 120 et 200 snaps sont respectivement mises en valeur par des pics des critères ϕ_{E_m}, ϕ_{R_c} et ϕ_{H_m} aux instants associés. Nous proposons ainsi de pallier ce problème en considérant conjointement plusieurs critères au cours de l'estimation des frontières structurelles.

Le tableau 6.2 référence les performances moyennes issues de l'union des frontières structurelles obtenues par le seuillage optimal des critères ϕ_{H_m}, ϕ_{R_c} et ϕ_{E_m} pour les tolérances de 0.5 s et 3 s. Cette méthode de fusion est clairement sous-optimale mais permet d'entrevoir les performances issues d'une combinaison de ces critères : cette union implique un nombre de frontières structurelles estimées très grand devant le nombre de frontières de référence, comme le montre le grand écart entre les valeurs moyennes de P_{br} et R_{br} pour les deux tolérances. En effet, si une même frontière est estimée par deux critères, alors celle-ci est comptée deux fois. Les valeurs de R_{br} pour tol=0.5 s et 3 s sont assez élevées en comparaison des valeurs oracles des critères séparés du tableau 6.1. $R_{br} = 78.52\%$ indique ainsi que la majeure partie des frontières de référence est contenue dans l'ensemble des frontières estimées pour tol=3 s.

Nous étudions dans cette partie le potentiel lié à la combinaison linéaire des trois meilleurs critères parmi ceux considérés dans la partie précédente. On se place de nouveau dans un cadre oracle.

6.1.3.1 Choix et combinaison des critères

Les résultats du tableau 6.1 nous permettent de sélectionner pour chacun des critères étudiés le type de descripteur maximisant leur F_{br} moyen sur RWC Pop. Nous considérons ainsi le critère de rupture d'homogénéité ϕ_{H_m} et de détection d'événements ϕ_{E_m} calculés sur les MFCC, et le critère de rupture de répétition ϕ_{R_c} calculé sur les vecteurs de chroma.

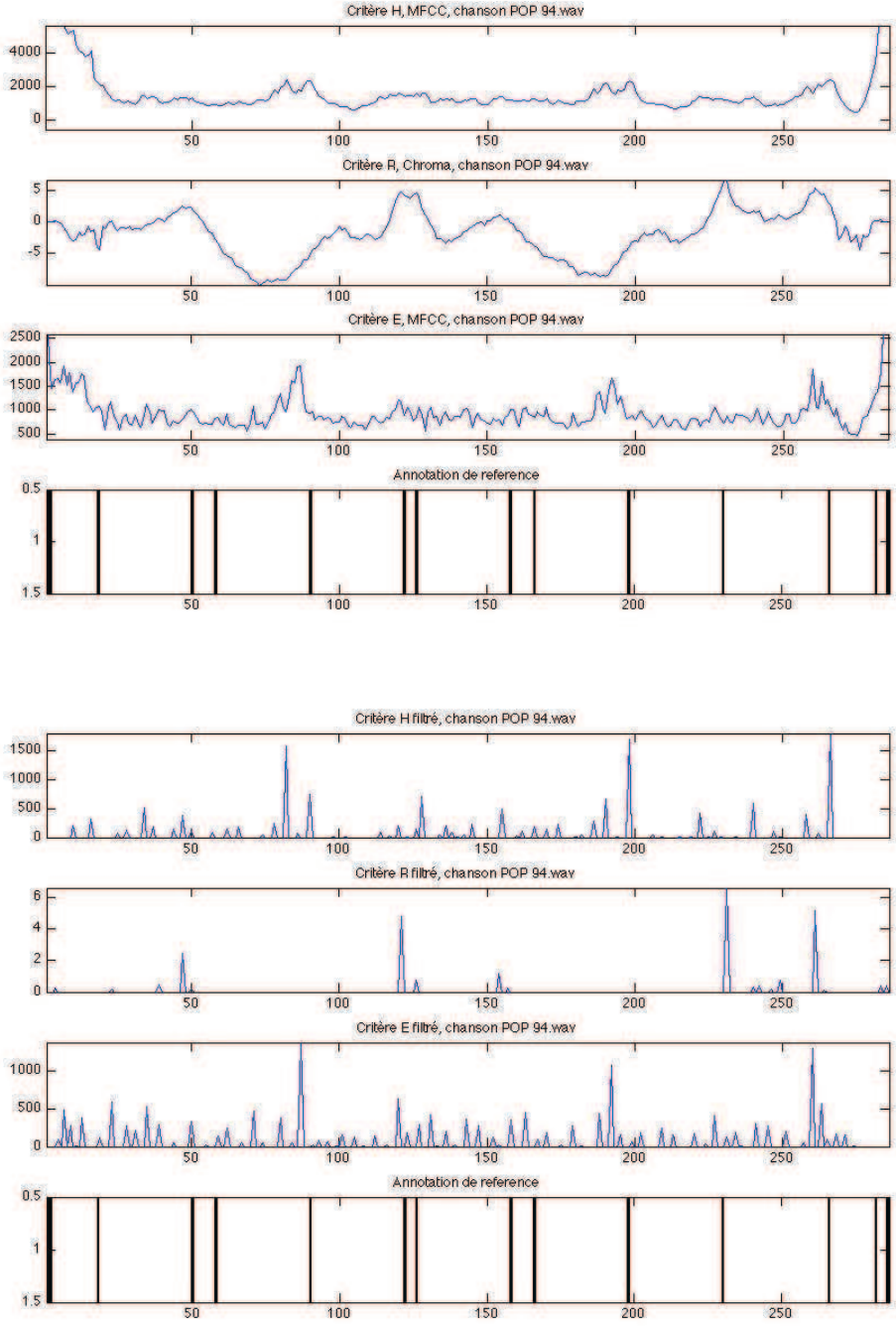

FIGURE 6.2 – Trois critères calculés pour le morceau numéro 94 de la base RWC Pop mis en regard des frontières structurelles de référence, sans filtrage (haut) puis filtrés (bas).

6.1.3.2 Protocole d'évaluation oracle

Nous calculons pour chaque morceau de la base de test (les 100 morceaux de RWC Pop) le critère combiné défini par la formule suivante :

$$\phi_{CL} = \lambda_1 \phi^f_{H_m} + \lambda_2 \phi^f_{R_c} + \lambda_3 \phi^f_{E_m} \tag{6.1}$$

où $\phi^f_{H_m}$, $\phi^f_{R_c}$ et $\phi^f_{E_m}$ correspondent aux versions filtrées et normalisées des critères ϕ_{H_m}, ϕ_{R_c}, et ϕ_{E_m} considérés dans la partie 4.1.5.1. Les frontières estimées sont obtenues en sélectionnant les pics de ϕ_{CL} dépassant le seuil optimal déterminé *a posteriori* à la manière de la partie 6.1.2.1. Les poids λ_1, λ_2 et λ_3 sont réglés afin de maximiser le F_{br} entre les frontières estimées et celles de référence pour chaque morceau en considérant la grille de valeurs telle que $\lambda_i \in [0, 1]$ et $\sum_{i=1}^{3} \lambda_i = 1$. Le pas de recherche considéré est 0.01.

6.1.3.3 Résultats

Les distributions des coefficients de pondération λ_1 et λ_2 qui maximisent le F_{br} pour chaque morceau de la base et pour les deux tolérances sont représentées dans la figure 6.3. Le troisième poids peut être déduit des deux autres par la formule $\lambda_3 = 1 - \lambda_1 - \lambda_2$. On observe que la grille de valeurs considérée pour λ_1 et λ_2 est globalement couverte et montre que plus d'un critère est utilisé afin d'obtenir les performances optimales. On note une augmentation de la concentration au voisinage de $\lambda = 1$, ce qui est cohérent avec le fait que ϕ_{H_m} soit le critère audio le plus efficace lorsqu'il est utilisé seul, comme le montre le tableau 6.3.

Le tableau 6.3 répertorie les performances oracles moyennes obtenues sur RWC Pop. On observe que les F_{br} moyennes sont meilleures dans le cas de l'approche multicritère : on observe une augmentation d'environ 2% pour tol=0.5 s et de 5% pour tol=3 s en comparaison de l'approche mono-critère. Les distributions des F_{br} oracles obtenus pour les différents morceaux de la base et pour les deux tolérances sont représentées dans la figure 6.4. On observe une distribution relativement étalée constituée de deux modes centrés en 30% et en 45% pour tol=0.5 s, et une distribution mono-modale plus concentrée autour de la valeur 70% pour tol=3 s.

Les histogrammes de la figure 6.5 représentent la différence des F_{br} oracles obtenus avec ϕ_{CL} par rapport à ϕ_{H_m} pour chaque morceau de la base et pour les deux tolérances considérées. Ces différences sont en général peu marquées, avec un grand nombre de morceaux associés à une différence de F_{br} inférieure à 2.5% : environ 70 morceaux pour tol=0.5 s et moins de 45 morceau pour tol=3 s. On note néanmoins une amélioration sensible pour les autres morceaux, jusqu'à environ 25% pour tol=3 s, qui rend l'approche multicritère intéressante face à la diversité des indices structurels des différents morceaux de musique.

Ces résultats préliminaires sont en faveur de la combinaison des critères afin d'obtenir de meilleures estimations des frontières structurelles. L'augmentation potentielle est relativement modeste malgré nos expériences dans un cadre oracle, ce qui peut être attribué à une combinaison assez "brutale" des critères par combinaison linéaire et du seuil de détection fixé pour chaque morceau de musique. Le recours à la formulation du problème de l'estimation des frontières par un processus d'optimisation va nous permettre d'"adoucir" le seuil de détection par l'introduction de contraintes structurelles, comme la contrainte e régularité à laquelle nous nous attachons dans la partie suivante.

Seuil de sélection des pics réglé pour tol=0.5 s			
	tol = 0.5 s		
	F_{br} (%)	P_{br} (%)	R_{br} (%)
ϕ_{CL}	**38.97**	47.36	36.64
ϕ_{H_m}	36.69	47.94	33.87
Seuil de sélection des pics réglé pour tol=3 s			
	tol = 3 s		
	F_{br} (%)	P_{br} (%)	R_{br} (%)
ϕ_{CL}	**69.13**	76.18	65.33
ϕ_{H_m}	64.19	71.81	60.70

TABLEAU 6.3 – Performances moyennes oracles sur RWC Pop lorsque l'on considère le critère ϕ_{CL} issu de la combinaison linéaire de ϕ_{H_m}, ϕ_{R_c} et ϕ_{E_m}. Les poids des critères et le seuil de sélection des pics sont réglés pour chaque morceau afin de maximiser son F_{br} pour tol=0.5 s (haut) et tol=3 s (bas). On rappelle dans les deux cas les performances oracles du critère ϕ_{H_m} obtenues dans la partie 6.1.2.2.

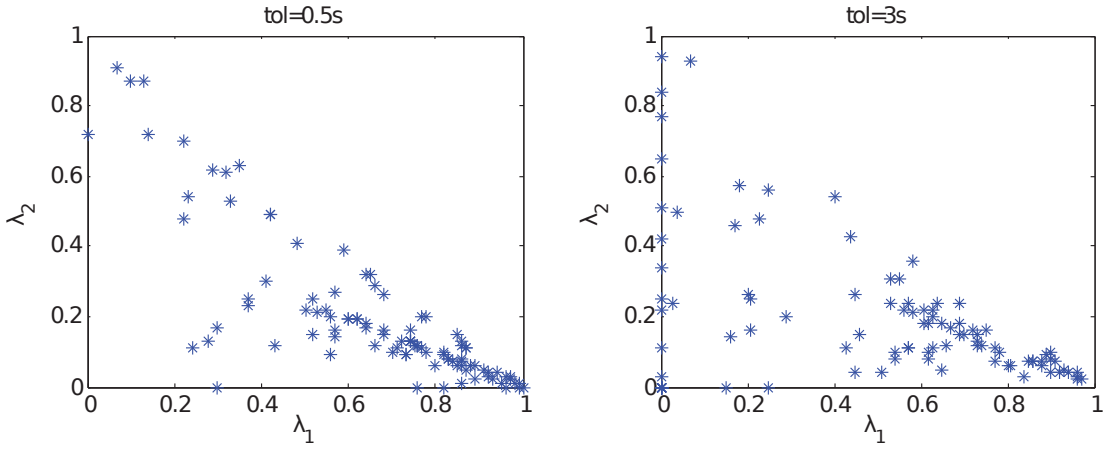

FIGURE 6.3 – Représentation des poids λ_1 et λ_2 optimaux ayant été obtenus pour les 100 morceaux de RWC Pop et pour les tolérances 0.5 s et 3 s.

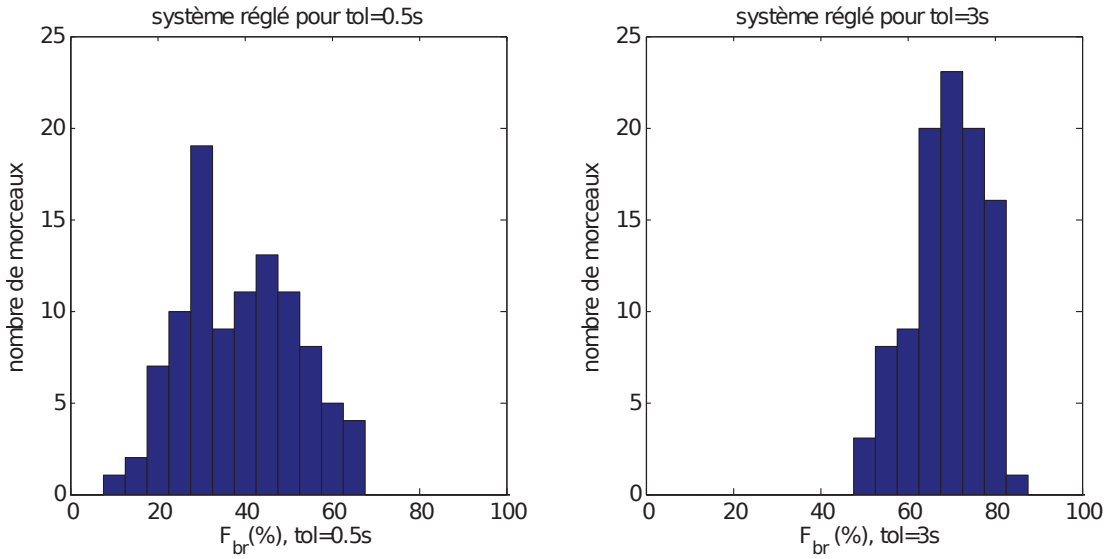

FIGURE 6.4 – Distribution des F_{br} oracles issus du réglage optimal des poids de la combinaison linéaire des critères ϕ_{H_m}, ϕ_{R_c}, et ϕ_{E_m} pour chaque chanson (tolérance considérée : 3 s).

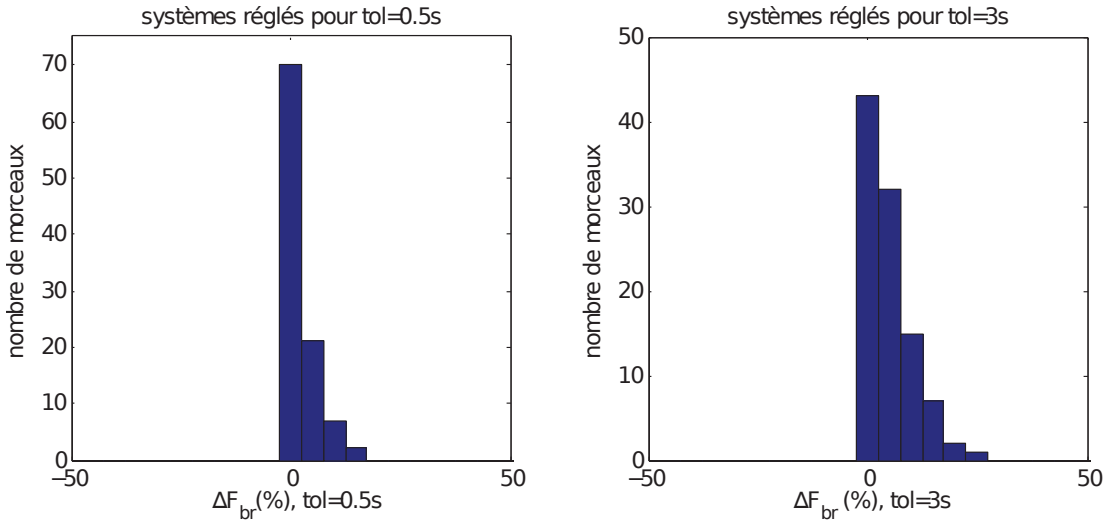

FIGURE 6.5 – Distribution des différences entre les F_{br} oracles obtenus dans le cas de ϕ_{CL} et de ϕ_{H_m} seul pour les différents morceaux de la base RWC Pop, et pour les tolérances de 0.5 s et 3 s.

6.2 Segmentation structurelle sous contrainte de régularité

Dans cette partie, nous considérons un système composé d'un seul critère audio et d'une contrainte de régularité structurelle afin d'étudier l'intérêt d'une telle contrainte dans le processus d'estimation des frontières structurelles. Nous nous plaçons sous l'hypothèse d'existence d'une seule pulsation structurelle, c'est-à-dire d'une taille préférentielle de bloc structurel.

6.2.1 Étude d'un modèle simple de contrainte de régularité

L'introduction d'une contrainte de régularité nous permet d'intégrer l'hypothèse de pulsation structurelle au processus d'estimation des frontières structurelles. Nous considérons ce problème de segmentation sous la forme du problème d'optimisation décrit au paragraphe 4.1.1. Ceci permet de s'affranchir du problème du réglage du seuil de sélection des pics des critères, et de travailler sur un système de segmentation opérationnel.

6.2.1.1 Système de segmentation étudié

Nous allons évaluer le système constitué d'un critère audio et du modèle de contrainte de régularité introduit dans la partie 4.1.3 combinés sous la forme d'un coût de segmentation. La recherche de la segmentation de coût minimal est réalisée à l'aide d'un algorithme de Viterbi (*cf.* partie 4.1.5.3).

Nous nous proposons de considérer le critère de rupture d'homogénéité calculé sur les MFCC, ϕ_{H_m}, ayant donné les performances les plus hautes lors de l'étude en oracle dans la partie 6.1.2.2. Ce critère est filtré par le critère de Seck comme décrit dans la partie 4.1.5.1 puis normalisé par rapport à sa valeur maximale.

Nous considérons un ensemble de fonctions de régularité non-convexes et convexes, issues de la famille de fonctions Ψ_α introduite dans la partie 4.1.3. Pour chaque chanson, la pulsation structurelle τ est fixée au nombre de temps musicaux le plus proche de 16 s parmi les valeurs 16, 32 et 64 temps. Ceci permet de garantir une pulsation structurelle de 16 snaps pour l'ensemble des morceaux étudiés qui est indépendante de leur tempo.

6.2.1.2 Protocoles expérimentaux

Nous étudions dans un premier temps l'impact du paramètre de convexité α ainsi que celui du paramètre de pondération λ contrôlant l'importance de la contrainte de régularité par rapport au critère audio considéré. Pour cela, le système complet est évalué sur la base RWC Pop et pour les tolérances de 0.5 s et 3 s pour un ensemble de réglages de ces paramètres : α varie entre 0 et 3 avec un pas de 0.1, ce qui permet de considérer un ensemble représentatif des comportements de la famille de fonctions considérée, λ varie entre 0 et 1 avec un pas de 0.01.

Nous menons dans un second temps une évaluation plus réaliste du système selon un processus de validation croisée (ou *jackknife* [Mil74]) sur RWC Pop. Il s'agit de régler notre système sur 80 morceaux de la base afin de le tester sur les 20 restants. Le couple (α, λ) est réglé de manière à maximiser le F_{br} moyen pour les tolérances de 0.5 s et 3 s sur les 80 morceaux qui constituent la base de développement et en considérant la même grille de valeurs que pour l'expérience précédente. On réitère ce processus en prenant 20 morceaux différents de la précédente base de test pour l'évaluation du

Étape de validation croisée	indices des morceaux constituant Test
1	6, 69, 5, 9, 54, 13, 83, 84, 73, 18, 66, 52, 98, 64, 80, 44, 42, 85, 10, 16
2	21, 39, 86, 79, 7, 40, 53, 43, 67, 62, 30, 45, 2, 99, 22, 14, 37, 25, 55, 35
3	96, 93, 8, 76, 31, 48, 59, 94, 46, 100, 32, 72, 68, 57, 74, 65, 19, 15, 97, 23
4	3, 58, 89, 70, 24, 36, 49, 95, 20, 87, 63, 34, 26, 47, 51, 12, 71, 28, 41, 75
5	29, 33, 78, 27, 90, 92, 81, 38, 61, 4, 91, 82, 77, 11, 56, 1, 50, 17, 60, 88

TABLEAU 6.4 – Indices des morceaux de la base RWC Pop pour chaque ensemble de test utilisé dans le processus de validation croisée. À chaque étape, l'ensemble d'apprentissage Dev correspond à l'ensemble complémentaire à Test par rapport à cette base.

système et les 80 morceaux restants comme base de développement. On procède ainsi jusqu'à avoir considéré les 100 morceaux de RWC dans les bases de test. On calcule enfin la moyenne des performances obtenues sur chacune des bases de test considérées. La répartition des morceaux de RWC Pop pour chaque ensemble de test s'est faite aléatoirement. Leurs indices dans le base sont spécifiés dans le tableau 6.4.

6.2.1.3 Résultats de l'étude des paramètres α et λ

Nous nous intéressons tout d'abord à l'impact du réglage de λ sur les performances moyennes du système considéré via les courbes de la figure 6.6. Chaque courbe caractérise l'évolution du F_{br} moyen sur RWC Pop en fonction de λ pour une valeur de α particulière et pour les tolérances de 0.5 s et 3 s. Pour chaque α, ce F_{br} décroit lorsque λ atteint ses valeurs extrêmes (0 ou 1). Ceci montre d'une part l'intérêt de considérer une contrainte de régularité pour estimer les frontières structurelles, et d'autre part qu'un poids trop fort sur cette contrainte aboutit à ignorer le critère et conduit à une segmentation trop régulière pour être pertinente. L'évolution de F_{br} en fonction de λ est relativement simple et passe par un maximum global. Elle a tendance à s'aplatir lorsque α augmente.

La figure 6.7 rassemble les courbes d'évolution du F_{br} moyen maximal en fonction de α, c'est-à-dire que λ est réglé de manière à maximiser le F_{br} moyen sur RWC Pop pour chaque α et pour les deux tolérances. Nous faisons apparaître à titre indicatif les valeurs des F_{br} oracles obtenues pour ϕ_{H_m} dans la partie 6.1.2.2 en pointillés. On montre ainsi que le système étudié obtient des résultats meilleurs que ceux du cas oracle, pour un bon réglage de λ : c'est le cas pour tout α lorsque l'on considère tol=3 s, et pour $\alpha \in [0.1, 0.8]$ dans le cas où tol=0.5 s. Enfin, ces courbes mettent en valeur deux des réglages optimaux du système : $(\alpha = 0.1, \lambda = 0.38)$ permet d'obtenir un F_{br} moyen de 38.70% pour tol=0.5 s et $(\alpha = 1.1, \lambda = 0.21)$ donne 68.68% pour tol=3 s.

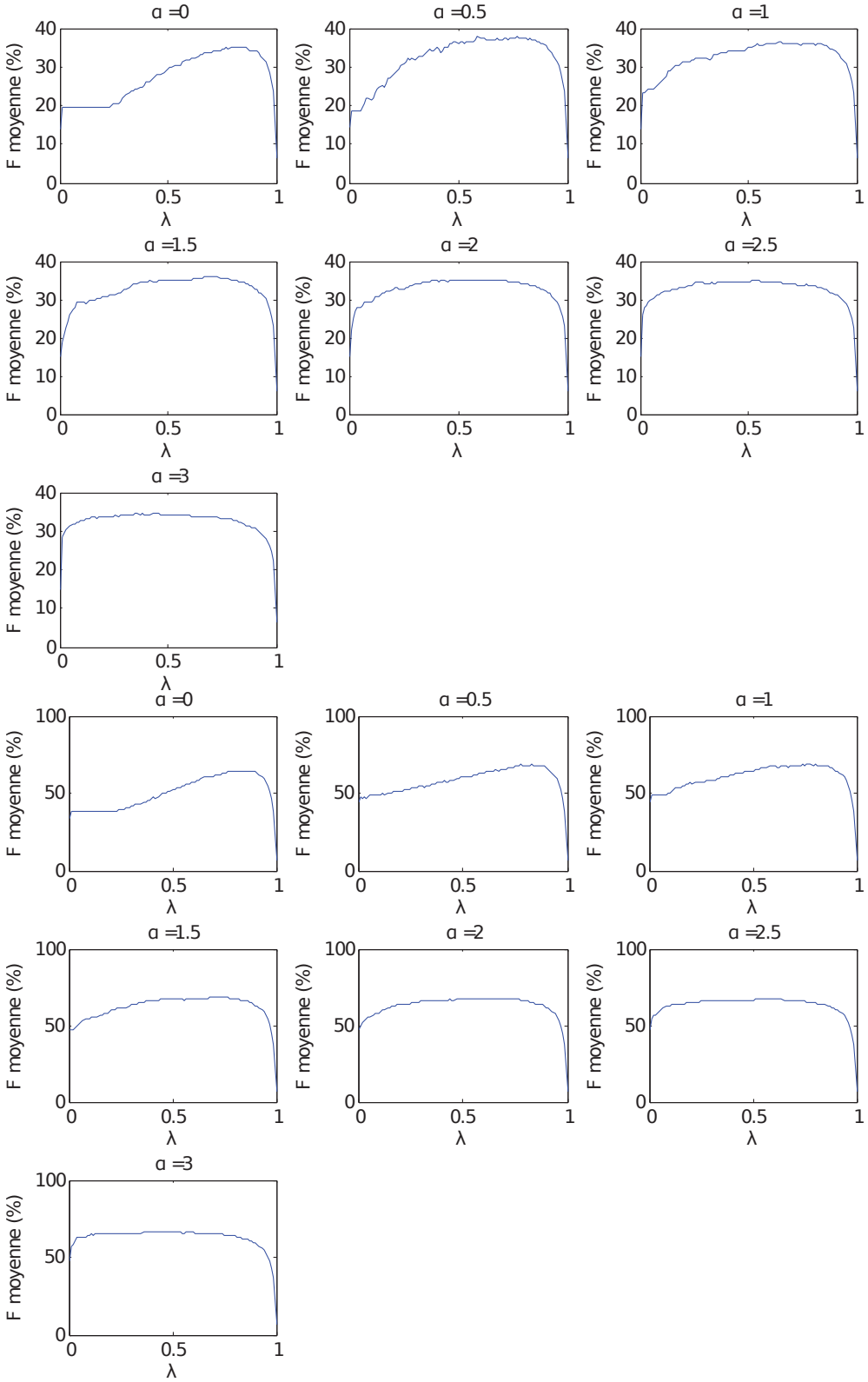

FIGURE 6.6 – Évolution du F_{br} moyen sur RWC Pop en fonction de λ et pour un sous-ensemble des valeurs de α considérées (α compris entre 0 et 3 avec un pas de 0.5). Les sept premières courbes correspondent à la tolérance de 0.5 s, les sept dernières à la tolérance de 3 s.

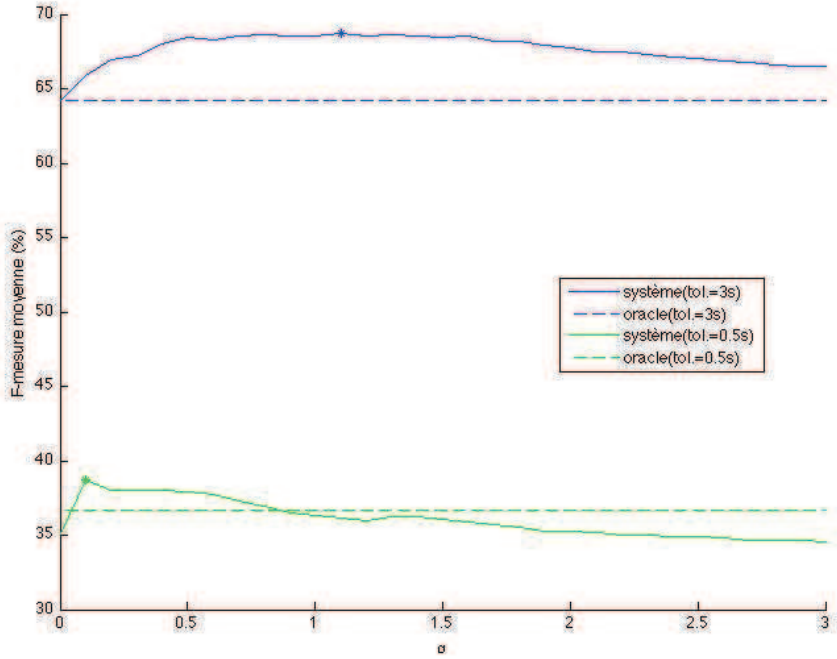

FIGURE 6.7 – Évolution du F_{br} moyen maximal sur RWC par rapport aux valeurs de α pour les tolérances de 0.5 s et 3 s. Dans le cas de la courbe bleue, on a réglé λ pour chaque α de manière à maximiser le F_{br} moyen sur RWC pour la tolérance de 3 s. Dans le cas de la courbe verte, il s'agit de la tolérance de 0.5 s.

Réglage des paramètres en maximisant le F_{br} moyen pour tol=0.5 s								
	paramètres		tol = 0.5 s			tol = 3 s		
	α	λ	F_{br} (%)	P_{br} (%)	R_{br} (%)	F_{br} (%)	P_{br} (%)	R_{br} (%)
Étape 1	0.1	0.33	38.36	42.63	36.34	69.85	73.08	67.91
Étape 2	0.1	0.24	42.17	43.78	42.21	67.09	69.76	66.94
Étape 3	0.1	0.38	35.60	36.60	35.48	60.13	62.15	59.56
Étape 4	0.1	0.38	36.48	40.51	33.42	56.70	62.78	52.13
Étape 5	0.1	0.38	32.97	36.21	31.20	57.92	63.08	55.25
Performances sur 1-100			**37.12**	39.95	35.49	**60.83**	65.18	58.41

Réglage des paramètres en maximisant le F_{br} moyen pour tol=3 s								
	paramètres		tol = 0.5 s			tol = 3 s		
	α	λ	F_{br} (%)	P_{br} (%)	R_{br} (%)	F_{br} (%)	P_{br} (%)	R_{br} (%)
Étape 1	1.1	0.21	37.92	40.29	36.34	69.85	73.08	67.91
Étape 2	0.8	0.21	39.09	39.70	39.73	65.70	66.49	67.00
Étape 3	1.1	0.21	34.74	34.09	36.56	67.04	65.59	70.62
Étape 4	0.7	0.16	35.33	37.28	33.95	67.36	70.67	65.24
Étape 5	1.1	0.21	32.70	33.70	33.04	69.52	71.68	69.92
Performances sur 1-100			**35.96**	37.01	35.92	**67.90**	69.50	68.14

TABLEAU 6.5 – Récapitulatif des performances moyennes obtenues dans le cadre du processus de validation croisée du système combinant le critère ϕ_{H_m} et la contrainte de régularité Ψ_α. Pour chaque étape de cette validation, α^* et λ^* sont issues du réglage de α et λ sur l'ensemble de développement "Dev". Le système utilise ces réglages pour être testé sur "Test" (la répartition des morceaux de RWC Pop est précisée dans le tableau 6.4). Le réglage est effectué de manière à maximiser le F_{br} moyen sur "Dev" pour tol=0.5 s (haut) et pour tol=3 s (bas).

6.2.1.4 Résultats de l'évaluation du système considéré par validation croisée

Le tableau 6.5 répertorie les performances moyennes obtenues pour les différentes étapes du processus de validation croisée en considérant les tolérances de 0.5 s et 3 s. Les F_{br} moyens obtenus pour les différentes étapes sont comparables lorsque le réglage de α et λ est effectué pour une tolérance de 3 s. On observe que le système tend à sous-segmenter les morceaux : P_{br} est en général supérieur à R_{br}. On remarque que le réglage optimal ($\alpha = 1.1, \lambda = 0.21$) est obtenu sur trois des cinq ensembles de développement utilisés. Lorsque le réglage s'effectue pour tol=0.5 s, les F_{br} moyens obtenus pour chaque étapes sont moins comparables, ce qui semble indiquer que les ensembles de test ne sont pas complètement homogènes. On observe cependant les mêmes tendances que pour le réglage considérant tol=3 s.

Les performances moyennes sur les cinq étapes permettent d'obtenir des performances supérieures à l'évaluation oracle du critère ϕ_{H_m} seul (*cf.* tableau 6.1) : on obtient $F_{br} = 67.90\% > 64.19\%$ pour tol=3 s et $F_{br} = 37.12\% > 36.69\%$ pour tol=0.5 s. Ceci montre que la prise en compte de notre contrainte de régularité améliore notre estimation des frontières structurelles sur RWC Pop.

6.2.2 Limites liées à la contrainte de régularité

Le modèle de contrainte de régularité considéré se fonde sur l'hypothèse d'existence d'une seule pulsation structurelle. Cependant, la structure de certains morceaux peut en présenter plusieurs. Nous étudions ici l'apport lié à l'utilisation de contraintes de régularité plus compliquées en considérant pour chaque morceau de RWC Pop le coût de régularité "idéal" directement issu de l'histogramme des tailles des segments des annotations de référence (*cf.* partie 4.1.4) [1]. Nous associons une taille nulle aux segments pour lesquels le nombre de snaps n'a pu être déterminé à la main. Les annotations peuvent faire apparaître des *tuilages* entre deux blocs structurels successifs de taille n et m, c'est-à-dire qu'ils se superposent sur une taille p (certaines propriétés de la fin du premier bloc correspondent au début du second) [BDSV11]. Dans le cadre du calcul des fonctions de régularité "idéales" nous considérons que le premier bloc se réalise entièrement et que le commencement du second est tronqué.

Nous évaluons le système de la partie 6.2.1.1 comprenant le critère de rupture d'homogénéité ϕ_{H_m} et la contrainte de régularité "idéale" propre à chaque morceau. Nous calculons cette fois-ci les performances moyennes obtenues sur RWC Pop en ayant réglé λ afin de maximiser le F_{br} pour chaque morceau, et pour les tolérances de 0.5 s et 3 s. Ces performances sont répertoriées dans le tableau 6.6. L'histogramme de la figure 6.8 rassemble les F_{br} obtenues pour les 100 morceaux de RWC. On observe que les quatre distributions sont globalement mono-modales. Celle des F_{br} pour tol=0.5 s sont plus étalées sur l'intervalle [0%,100%] que pour tol=3 s qui se concentre davantage sur l'intervalle [50%,100%].

L'apport lié à l'utilisation de la contrainte de régularité "idéale" peut être visualisé à l'aide des histogrammes de la figure 6.9. Ces deux histogrammes permettent d'observer la distribution des différences entre les F_{br} du système abrégé *régularité "idéale'*, combinant le critère ϕ_{H_m} avec la contrainte idéale, et ceux du système de référence décrit dans la partie 6.2.1.1 avec $\tau = 16$ snaps, $\alpha = 0.5$, $\lambda = 0.17$, pour chacun des morceaux de la base et pour les deux tolérances. Le paramètre λ est réglé pour tol=0.5 s dans le cas de l'histogramme de gauche, et pour tol=3 s pour celui de droite. L'histogramme de gauche met en valeur le fait que le système *régularité "idéale" (0.5s)* est meilleur que le système de référence sur seulement 61 morceaux sur les 100 de la base pour tol=0.5 s. De la même manière, le système *régularité "idéale" (3s)* est meilleur que le système de référence sur seulement 44 morceaux de la base pour tol=3 s. Pour les autres morceaux, le système contenant le coût de régularité $\Psi_{\alpha=0.5,\tau=16}$ donne de meilleurs performances.

Les histogrammes de la figure 6.10 permettent de visualiser la distribution des valeurs de λ qui maximisent les F_{br} pour chaque morceau et pour les deux tolérances. On remarque qu'elles tendent à se concentrer autour de 0 et 1 ce qui implique de manière assez surprenante que la contrainte de régularité "idéale" est en général assez peu considérée (un seul morceau n'utilise pas du tout la contrainte de régularité sur l'ensemble des tolérances considérées). On note que la configuration optimale du système pour une dizaine de morceaux pour tol=0.5 s et 3 s implique que seule la contrainte est utilisée : on a $\lambda = 1$ pour 6 morceaux quand tol=0.5 s et 8 morceaux quand tol=3 s.

1. La version des annotations de RWC Pop que nous utilisons est telle qu'à chaque bloc structurel annoté est attribuée une étiquette décrivant le nombre de snaps qu'il contient. Pour des raisons de cohérence avec les parties précédentes, nous convertissons ces tailles en nombre de temps. On considère que l'échelle des snaps correspond à une échelle proportionnelle à celle des temps musicaux, synchrone à l'échelle associée aux premiers temps des mesures musicales, et dont la période entre deux unités successives est proche de 1 s.

	tol = 0.5 s			tol = 3 s		
	$F_{br}(\%)$	$P_{br}(\%)$	$R_{br}(\%)$	$F_{br}(\%)$	$P_{br}(\%)$	$R_{br}(\%)$
référence	37.25	38.46	37.03	68.42	70.19	68.42
régularité “idéale” (0.5 s)	**48.60**	52.61	46.89	63.68	68.90	61.32
régularité “idéale” (3 s)	39.20	42.21	37.31	**70.65**	76.38	67.26

TABLEAU 6.6 – Performances moyennes obtenues pour les tolérances de 0.5 et 3 s sur RWC Pop pour les systèmes *référence* (contrainte de régularité $\Psi_{\tau=16,\alpha=0.5}$ avec $\lambda = 0.17$), *régularité “idéale” (0.5 s)* (contrainte de régularité “idéale” et λ réglé pour chaque chanson afin de maximiser la F-mesure à 0.5 s), et *régularité “idéale” (3 s)* (contrainte de régularité “idéale” et λ réglé pour chaque chanson afin de maximiser la F-mesure à 3 s).

Ces observations tendent à favoriser l’usage du modèle simple de régularité proposé par rapport à des contraintes complexes.

6.3 Segmentation structurelle par analyse multicritère et contrainte de régularité

Nous allons maintenant étudier un système d’estimation des frontières structurelles combinant trois critères audio et une contrainte de régularité par l’intermédiaire de l’optimisation de coût décrite dans la partie 4.1.1. Cette optimisation permet de traiter le problème de la sélection les pics des critères combinés évoqué dans la partie 6.1.3.3 par un processus opérationnel. Le système est évalué à l’aide d’une procédure de validation croisée.

6.3.1 Système considéré

Le système utilise les critères de rupture $\phi^f_{H_m}$, $\phi^f_{R_c}$ et le critère de détection d’événements $\phi^f_{E_m}$ qui sont respectivement pondérés par les paramètres λ_1, λ_2, λ_3 et combinés en un critère ϕ_{CL} selon l’équation 6.1. On en déduit le coût Φ_{CL} qui associe à tout bloc s_k la somme des valeurs prises par ϕ_{CL} entre ses instants de début et de fin à la manière du système IRISA10 de la partie 5.2.1. La contrainte de régularité structurelle Ψ_α est celle introduite dans la partie 4.1.3. La pulsation structurelle considérée est $\tau = 16$ snaps : les critères étant exprimés à l’échelle des temps musicaux, on choisit ainsi le nombre de temps le plus proche de 16 s parmi les valeurs 8, 16, 32, 64 et 128 temps.

Les critères audio et la contrainte de régularité sont ensuite combinés sous la forme d’un coût de segmentation C par la formule suivante :

$$C = \Phi_{CL} + \lambda_4 \Psi_\alpha \tag{6.2}$$

6.3.2 Protocole expérimental

Nous utilisons la procédure de validation croisée et les divisions de RWC Pop en bases de développement et de test utilisées dans la partie 6.2.1.2. Les paramètres réglés sur la base de développement sont maintenant λ_1, λ_2, λ_3, λ_4 et α. Les poids λ_i sont

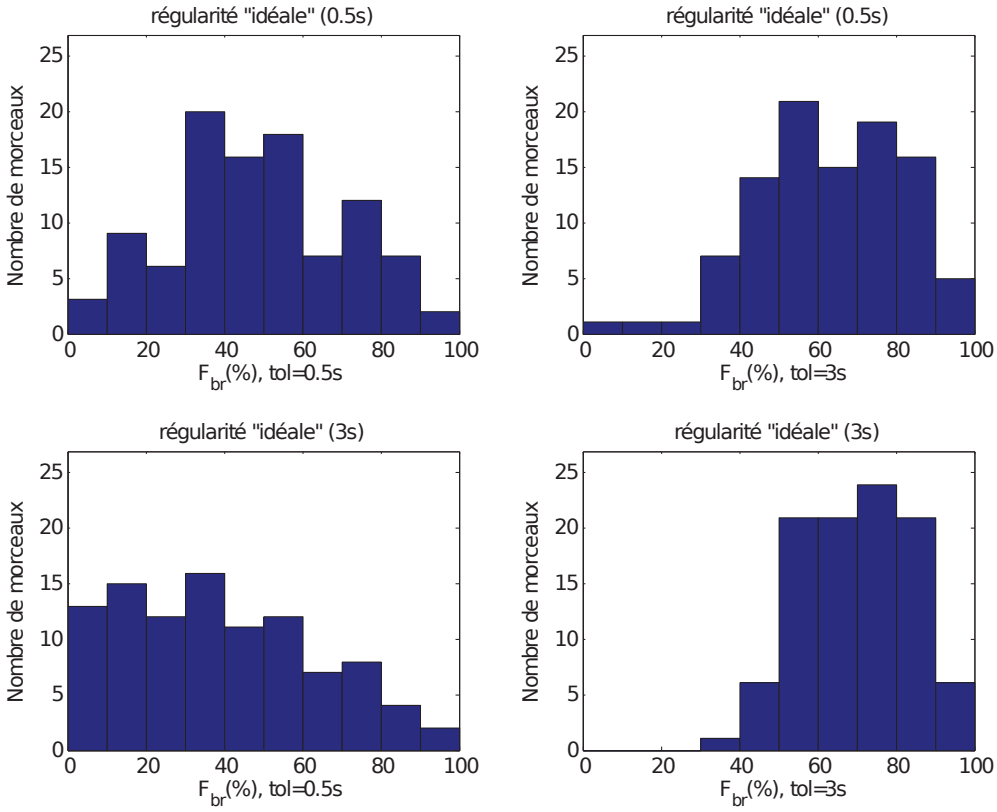

FIGURE 6.8 – Distribution des valeurs de F-mesures obtenues pour les 100 morceaux de RWC Pop lorsque λ est réglé de manière à maximiser F_{br} à 0.5 s (*régularité "idéale" (0.5 s)*) ou celle à 3 s (*régularité "idéale" (3 s)*).

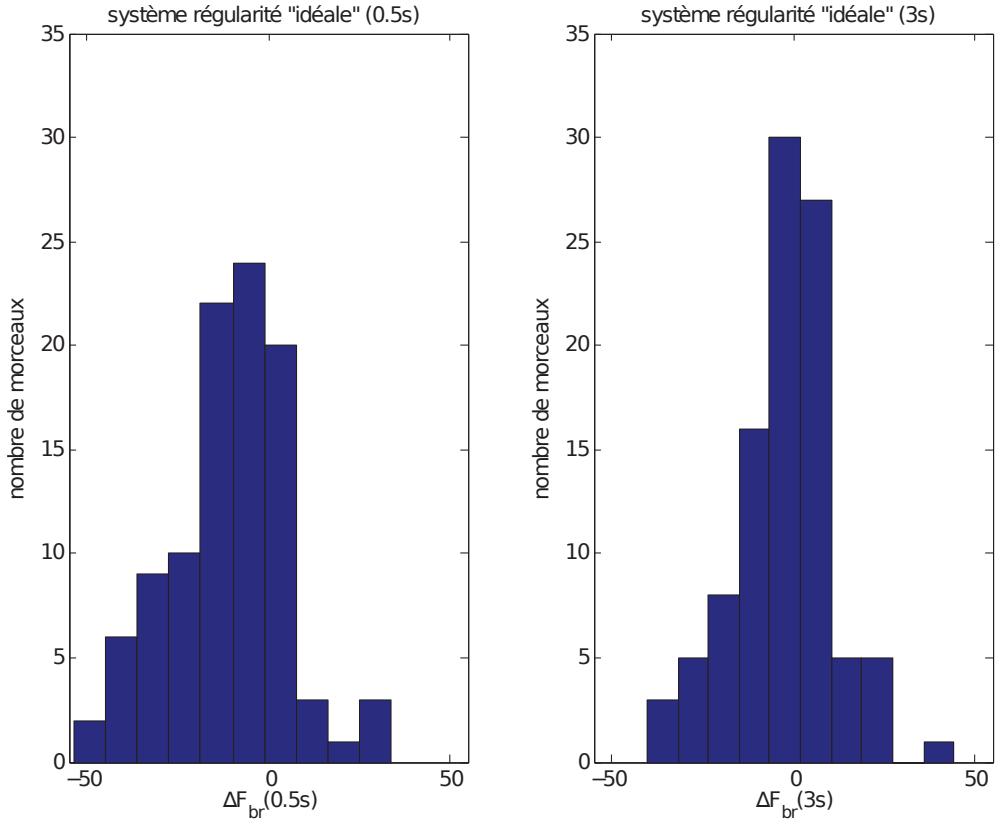

FIGURE 6.9 – Distribution des différences entre les F_{br} issues des système *régularité "idéale" (0.5 s)* et *régularité "idéale" (3 s)* par rapport à celles du système de référence décrit dans la partie 6.2.1.1 avec $\tau = 16$ snaps, $\alpha = 0.5$, $\lambda = 0.17$, pour chaque morceau de RWC Pop et pour la tolérance à 3 s.

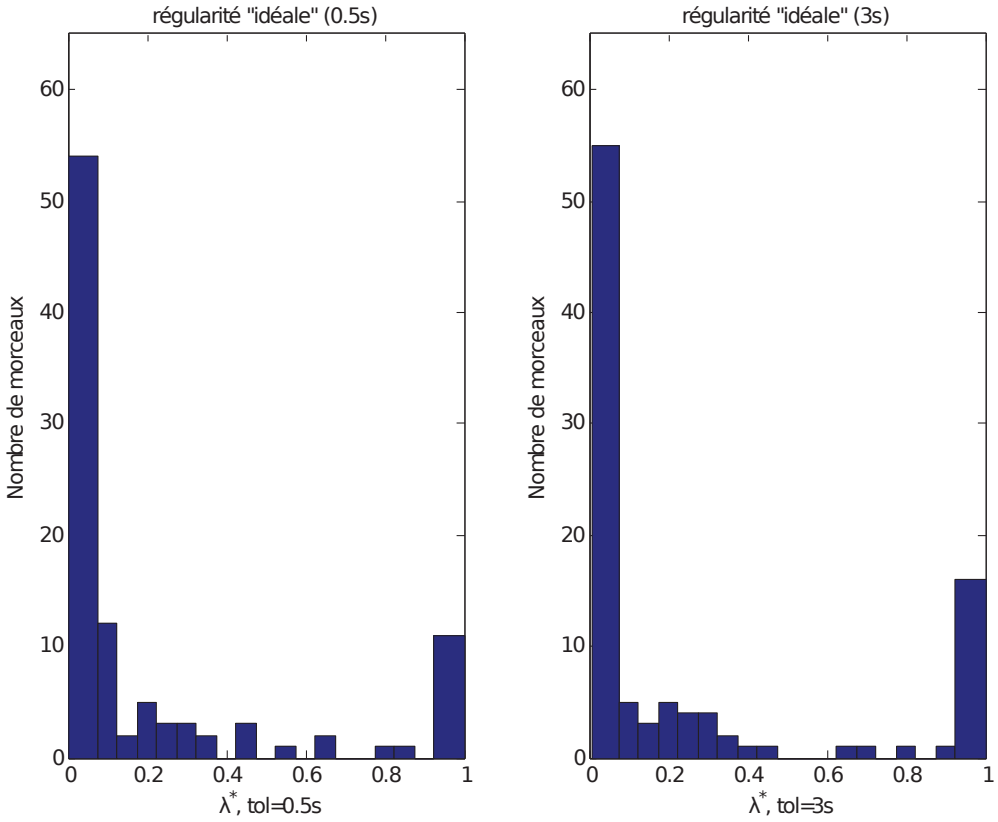

FIGURE 6.10 – Distribution des valeurs de λ maximisant le F_{br} des morceaux de RWC Pop pour une tolérance de 0.5 s (gauche) et une tolérance de 3 s (droite).

Réglage des paramètres en maximisant le F_{br} moyen pour tol=0.5 s											
	paramètres					tol = 0.5 s			tol = 3 s		
	α^*	λ_1^*	λ_2^*	λ_3^*	λ_4^*	F_{br} (%)	P_{br} (%)	R_{br} (%)	F_{br} (%)	P_{br} (%)	R_{br} (%)
Étape 1	0.1	0.6	0.1	0	0.3	39.07	43.44	35.77	63.51	69.35	59.39
Étape 2	0.2	0.6	0.1	0	0.3	42.67	45.08	41.92	63.00	66.61	61.71
Étape 3	0.1	0.7	0	0	0.3	34.57	34.91	35.07	60.82	61.74	61.27
Étape 4	0.1	0.6	0.1	0	0.3	36.54	40.68	33.41	59.70	66.41	54.69
Étape 5	0.1	0.6	0.1	0	0.3	31.98	35.15	30.23	65.85	70.93	63.48
Performances sur 1-100						**36.97**	39.85	35.28	**62.58**	67.01	60.11

Réglage des paramètres en maximisant le F_{br} moyen pour tol=3 s											
	paramètres					tol = 0.5 s			tol = 3 s		
	α^*	λ_1^*	λ_2^*	λ_3^*	λ_4^*	F_{br} (%)	P_{br} (%)	R_{br} (%)	F_{br} (%)	P_{br} (%)	R_{br} (%)
Étape 1	1.4	0.5	0	0.2	0.3	37.01	40.09	34.82	70.26	74.87	67.29
Étape 2	1.0	0.5	0	0.2	0.3	38.38	40.39	37.80	65.95	69.03	65.31
Étape 3	0.9	0.6	0	0.2	0.2	34.04	32.92	36.13	67.67	65.47	71.65
Étape 4	1.1	0.5	0	0.2	0.3	35.94	39.74	32.96	70.66	78.11	64.87
Étape 5	1.1	0.5	0	0.2	0.3	30.39	32.05	29.82	67.65	71.85	65.96
Performances sur 1-100						**35.15**	37.04	34.31	**68.44**	71.87	67.02

TABLEAU 6.7 – Récapitulatif des performances moyennes obtenues dans le cadre du processus de validation croisée du système combinant les critères ϕ_{H_m}, ϕ_{R_c}, ϕ_{E_m} et la contrainte de régularité Ψ_α. Pour chaque étape de cette validation, les cinq paramètres sont réglés sur l'ensemble de développement "Dev" avec une précision de 0.1. Le système utilise ces réglages pour être testé sur "Test" (la répartition des morceaux de RWC Pop est précisée dans le tableau 6.4). Le réglage est effectué de manière à maximiser le F_{br} moyen sur "Dev" pour tol=0.5 s (haut) et pour tol=3 s (bas).

considérés sur l'intervalle $[0, 1]$ avec un pas de 0.01, tels que $\sum_{i=1}^{4} \lambda_i = 1$, et le paramètre de convexité α est considéré sur l'intervalle $[0, 3]$ avec un pas de 0.1.

6.3.3 Résultats

Le tableau 6.7 répertorie les performances moyennes du système considéré pour les différentes étapes du processus de validation croisée. Les paramètres sont réglés de manière à maximiser le F_{br} moyen à chaque étape pour les tolérances tol=0.5 s et 3 s.

Les paramètres optimaux issus du réglage du système sur chaque base de développement montrent la prise en compte d'un critère audio principal, d'un critère audio secondaire et de la contrainte de régularité. Il s'agit des critères ϕ_{H_m} et ϕ_{R_c} pour tol=0.5 s, et de ϕ_{H_m} et ϕ_{E_m} pour tol=3 s. Le critère ϕ_{H_m} a un poids supérieur aux autres, ce qui est cohérent avec ses performances oracles présentés dans la partie 6.1.3.3 : λ_1^* est de l'ordre de 0.6 tandis que λ_2^* et λ_3^* sont de l'ordre de 0.2. De même, l'utilisation du critère ϕ_{R_c} devant ϕ_{E_m} pour la tolérance de 0.5 s et l'observation de la tendance inverse pour tol=3 s est conforme à leur classement selon leurs performances oracles (*cf.* tableau 6.1.2.2). On explique ces tendances par le fait que des événements courts peuvent être détectés à proximité des frontières des blocs structurels mais pas forcément exactement à la frontière.

Réglage des paramètres en maximisant le F_{br} moyen pour tol=0.5 s						
Systèmes	tol = 0.5 s			tol = 3 s		
	F_{br} (%)	P_{br} (%)	R_{br} (%)	F_{br} (%)	P_{br} (%)	R_{br} (%)
$(\phi_{H_m}, \Psi_\alpha)$	**37.12**	39.95	35.49	60.83	65.18	58.41
$(\phi_{H_m}, \phi_{R_c}, \phi_{E_m}, \Psi_\alpha)$	36.97	39.85	35.28	62.58	67.01	60.11
Réglage des paramètres en maximisant le F_{br} moyen pour tol=3 s						
Systèmes	tol = 0.5 s			tol = 3 s		
	F_{br} (%)	P_{br} (%)	R_{br} (%)	F_{br} (%)	P_{br} (%)	R_{br} (%)
$(\phi_{H_m}, \Psi_\alpha)$	35.96	37.01	35.92	67.90	69.50	68.14
$(\phi_{H_m}, \phi_{R_c}, \phi_{E_m}, \Psi_\alpha)$	35.15	37.04	34.31	**68.44**	71.87	67.02

TABLEAU 6.8 – Performances moyennes sur les différentes étapes du processus de validation croisée obtenues pour le système mono-critère avec une contrainte de régularité $(\phi_{H_m}, \Psi_\alpha)$ et celles du système multicritère avec la même contrainte de régularité $(\phi_{H_m}, \phi_{R_c}, \phi_{E_m}, \Psi_\alpha)$.

Les valeurs du paramètre de convexité obtenues pour les tolérances de 0.5 s et 3 s restent ainsi proches de celles issues de l'optimisation du système combinant ϕ_{H_m} et Ψ_α lors de la partie 6.2. De manière générale, les valeurs des paramètres apprises au cours de la validation croisée varient peu.

Le tableau 6.8 permet de mettre en regard les performances moyennes de la validation croisée du système considéré sur RWC Pop avec celles du système mono-critère constitué de la même contrainte de régularité qui ont été obtenues dans la partie 6.2. On observe ainsi que la prise en compte de plusieurs critères permet d'obtenir de meilleurs résultats lorsque les systèmes sont réglés sur les bases de développement afin de maximiser le F_{br} pour tol=3 s. On constate que ce n'est pas le cas lorsque la maximisation se fait pour tol=0.5 s, ce qui semble dû à la résolution de la grille de recherche des λ_i, trop grossière, en particulier pour l'étape 3 du processus de validation croisée. Un réglage plus fin des paramètres λ_i est à effectuer à l'avenir. Le calcul des intervalles de confiance sur la différence de performance du système multicritère par rapport au système mono-critère ne permet pas d'observer de différence significative entre eux : on obtient -0.15 ± 1.39 pour tol= 0.5 s quand les systèmes sont réglés pour cette tolérance, et on obtient 0.54 ± 1.62 pour tol=3 s en ayant réglé les systèmes selon cette même tolérance. Il n'est donc pas avantageux de considérer les deux critères supplémentaires ϕ_{R_c} et ϕ_{E_m} lorsque l'on utilise déjà la contrainte de régularité Ψ_α.

6.4 Étiquetage sémiotique

Nous nous intéressons dans cette partie à l'évaluation de plusieurs versions du système d'étiquetage présenté dans la partie 4.2. Nous supposons que les frontières structurelles de référence sont connues. Dans un premier temps, nous associons à chaque segment structurel la séquence de descripteurs contenue entre ses frontières. Le système est évalué pour les deux critères de sélection de modèle par automate considérés. Dans un deuxième temps, nous réévaluons le système en associant cette fois à chaque segment structurel les trois quarts de la séquence de descripteurs qu'il contient, lorsque sa taille est supérieure ou égale à 16 snaps. Ceci permet de considérer une portion des segments

	longueur de séquence	critère de sélection de modèle
Système 1	complète	AIC
Système 2	complète	CAA
Système 3	trois-quarts	AIC
Système 4	trois-quarts	CAA

TABLEAU 6.9 – Caractéristiques des versions du système d'étiquetage sémiotique évaluées. On précise si le système associe la totalité ou les trois quarts de la séquence de descripteurs associée à chaque segment structurel, et quel critère de sélection de modèle est utilisé entre le critère d'Information d'Akaike AIC et le critère auto-adaptatif CAA.

structurels *a priori* associée à la partie perceptible du système porteur carré décrit dans le cadre de la partie 3.5.3 dans l'optique d'attribuer la même étiquette sémiotique aux segments de même système porteur. Soulignons que l'on utilise ainsi une conséquence approchée du modèle système - contraste, mais pas le modèle en lui-même dans le processus d'étiquetage. La nature exploratoire du travail présenté ici implique que nous ne nous référerons pas aux performances d'autres méthodes de l'état de l'art.

6.4.1 Protocole expérimental

L'étiquetage sémiotique de la base RWC Pop étant en cours de finalisation, nous considérons pour nos évaluations un ensemble de 100 morceaux de musique annotés par l'IRCAM dans le cadre des évaluations Quaero. Il s'agit des bases Test09 et Test10 auxquels on ajoute les 6 premiers morceaux de la base Dev09. Ces bases ont été annotées selon une méthodologie différente de celle du chapitre 3 et sont décrites dans la partie 5.1.1.

Afin de distinguer les différentes versions du système utilisé, nous utiliserons les notations référencées dans le tableau 6.9. Ces systèmes sont évalués à l'aide des mesures de précision, rappel et F-mesure pour les groupements dyadiques de trames, respectivement notées pP, pR et pF et présentées dans la partie 5.1.2.2.

Les systèmes 1 et 3 nécessitent le réglage d'un paramètre de pondération a_{AIC} lié à la version du critère AIC considérée. a_{AIC} est fixé de manière à maximiser le pF moyen sur l'ensemble de développement parmi un ensemble de valeurs entre 0 et 10 avec un pas de 0.1. Ces systèmes sont évalués selon le processus de validation croisée décrit dans la partie 6.2.1.2. La répartition des morceaux de la base pour le test et le développement des systèmes selon les différentes étapes de la validation est effectuée aléatoirement. Elle est référencée dans le tableau 6.10 où chaque indice correspond au numéro de référence d'un morceau attribué dans le cadre de Quaero. L'association entre les morceaux et leur numéro de référence est précisée en annexe. Les critères de sélection des systèmes 2 et 4 ne nécessitent pas de réglage de paramètre. On calculera ainsi leurs performances moyennes sur l'ensemble de la base.

6.4.2 Résultats

La figure 6.11 représente les courbes d'évolution du pF moyen en fonction de a_{AIC}obtenues par les Systèmes 1 et 3 sur l' ensemble de développement de chaque étape de la validation croisée. On observe que chacune des courbes croît puis décroît

Étape de validation croisée	indices des morceaux constituant Test au sein de la base Quaero considérée
1	0010, 0174, 0009, 0014, 0115, 0018, 0226, 0229, 0208, 0034, 0165, 0107, 0293, 0158, 0219, 0091, 0089, 0230, 0015, 0028
2	0049, 0080, 0232, 0218, 0011, 0084, 0112, 0090, 0167, 0150, 0063, 0093, 0004, 0301, 0054, 0021, 0078, 0057, 0119, 0075
3	0281, 0275, 0012, 0215, 0066, 0098, 0138, 0276, 0094, 0302, 0067, 0186, 0168, 0127, 0211, 0163, 0038, 0024, 0282, 0055
4	0005, 0133, 0254, 0183, 0056, 0076, 0100, 0280, 0042, 0233, 0154, 0073, 0059, 0095, 0105, 0017, 0185, 0061, 0086, 0214
5	0062, 0070, 0217, 0060, 0261, 0270, 0223, 0079, 0145, 0007, 0262, 0224, 0216, 0016, 0123, 0002, 0104, 0030, 0140, 0253

TABLEAU 6.10 – Indices des morceaux de la base Quaero considérée pour chaque ensemble de test utilisé dans le processus de validation croisée. À chaque étape, l'ensemble d'apprentissage Dev correspond à l'ensemble complémentaire à Test par rapport à cette base.

sensiblement lorsque la valeur de a_{AIC} augmente, ce qui tend à montrer que l'intervalle considéré pour le réglage de ce paramètre est significatif.

Les performances moyennes obtenues pour les quatre systèmes considérés sont répertoriées dans le tableau 6.11. On observe que l'utilisation de l'un ou l'autre des critères permet d'aboutir à des résultats comparables, que l'on considère la totalité ou les trois quarts de la séquence de descripteurs associée aux segments structurels. Les performances obtenues avec le critère AIC sont en moyenne légèrement supérieures de celles du critère auto-adaptatif CAA. Cependant, le calcul de l'intervalle de confiance à 95% sur la différence de performance du Système 1 par rapport à celles du Système 2 est de $3.41 \pm 4.18\%$, et celui entre les performance du Système 3 par rapport au Système 4 est de $2.66 \pm 3.87\%$. Ceci implique qu'il n'y a pas de différence significative sur l'utilisation du critère AIC devant CAA. Notons que CAA reste avantageux par le fait qu'il ne comporte pas de paramètre à régler.

La comparaison des performances du Système 1 par rapport au Système 3 et du Système 2 par rapport au Système 4 permet de s'intéresser à l'apport lié à la prise en compte des trois quarts de la séquence des segments structurels de taille supérieure ou égale à 16 snaps. On observe une légère augmentation de l'ordre de 1% entre les systèmes 1 et 3, et de l'ordre de 2% entre les systèmes 2 et 4. On calcule un intervalle de confiance à 95% de $-1.06 \pm 1.59\%$ sur la différence de performance du Système 1 par rapport au Système 3 et de $-1.81 \pm 1.47\%$ sur la différence de performance du Système 2 par rapport au Système 4. Seul le second cas présente une différence significative : le Système 4 est meilleur que le Système 2. Ceci permet d'entrevoir l'intérêt de considérer la partie perceptible des systèmes porteurs dans l'optique de l'étiquetage sémiotique des segments structurels. Soulignons cependant que les annotations structurelles utilisées ne sont *a priori* pas basées sur le modèle système - contraste.

Ainsi, d'une part la distance entre les branches d'automate basée sur une mesure de probabilité est *a priori* peu robuste aux problèmes d'estimation des frontières structurelles ainsi qu'aux distorsions d'ordre temporel comme les ajouts et les troncatures. D'autre part, il nous semble prometteur de développer à l'avenir des systèmes

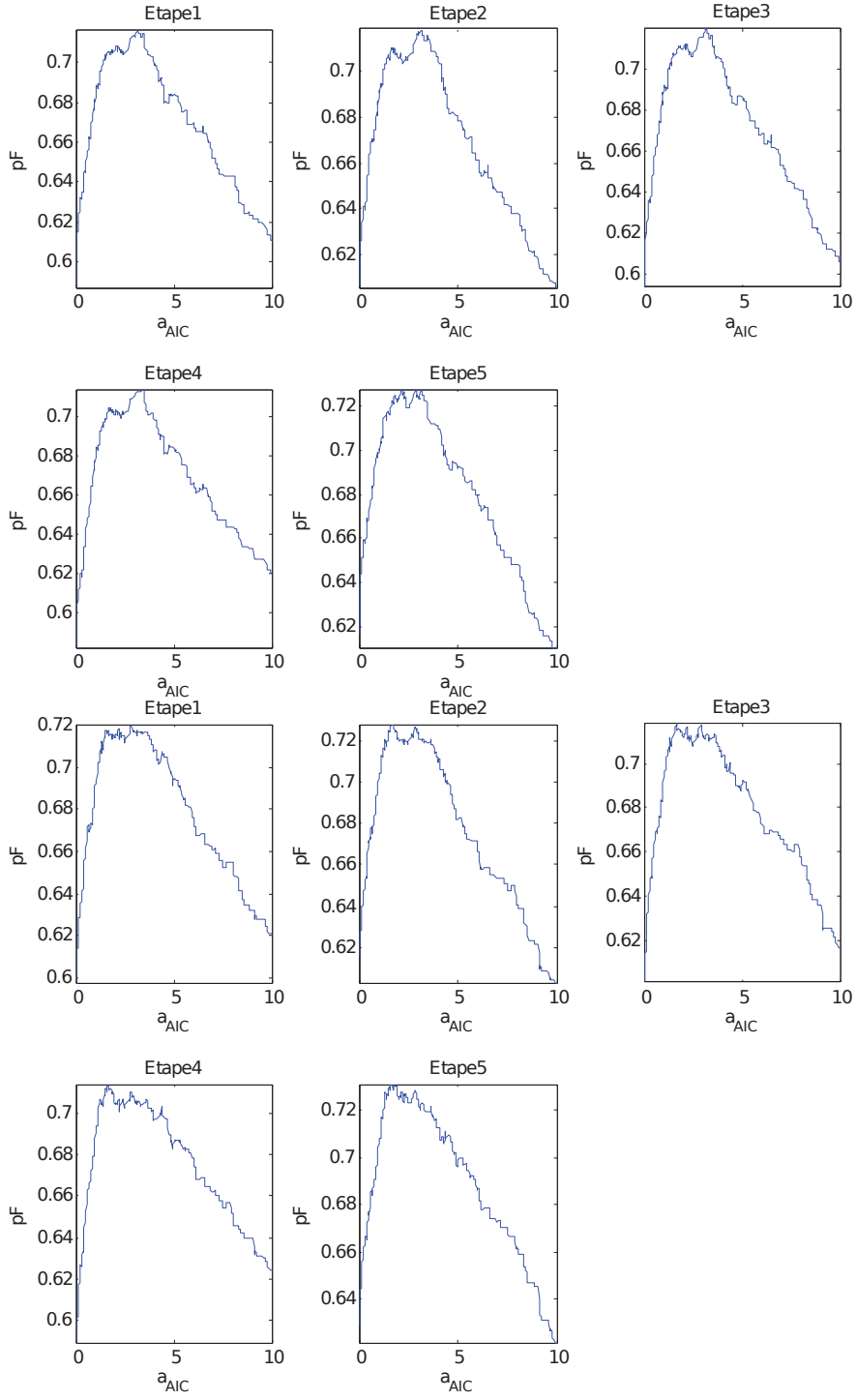

FIGURE 6.11 – Courbes d'évolution du pF moyen sur l'ensemble de développement pour chaque étape de la validation croisée, en fonction du paramètre a_{AIC}. Les cinq premières courbes concernent le Système 1, les cinq dernières concernent le Système 3.

Système 1				
	a^*_{AIC}	pF	pP	pR
Étape 1	3.18	72.76	79.38	72.17
Étape 2	3.18	72.15	77.55	74.69
Étape 3	3.18	71.47	68.26	80.55
Étape 4	3.44	72.71	67.22	86.30
Étape 5	2.16	55.63	99.00	41.02
Performances moyennes sur 1-100		**70.71**	74.52	75.37
Système 2				
Performances moyennes sur 1-100		**67.30**	66.69	73.32
Système 3				
	a^*_{AIC}	pF	pP	pR
Étape 1	2.81	72.01	77.09	72.64
Étape 2	1.71	69.66	85.25	64.25
Étape 3	1.71	73.72	79.13	73.60
Étape 4	1.64	75.16	80.39	75.89
Étape 5	1.71	68.31	84.95	63.17
Performances moyennes sur 1-100		**71.77**	81.36	69.91
Système 4				
Performances moyennes sur 1-100		**69.11**	66.72	76.63

TABLEAU 6.11 – Récapitulatif des performances moyennes obtenues dans le cadre du processus de validation croisée du système combinant les critères ϕ_{H_m}, ϕ_{R_c}, ϕ_{E_m} et la contrainte de régularité Ψ_α. Pour chaque étape de cette validation, a_{AIC} est réglé sur l'ensemble de développement "Dev" de manière à maximiser le pF moyen. Le système utilise ces réglages pour être testé sur "Test".

d'étiquetage sémiotique intégrant ce modèle d'organisation interne des segments structurels.

6.5 Résumé du chapitre

Ce chapitre présente dans un premier temps un ensemble d'évaluations permettant d'étudier une approche multicritère et l'utilisation d'une contrainte de régularité pour l'estimation des frontières structurelles. Ces évaluations sont réalisées sur la base constituée des 100 morceaux de RWC Pop avec les annotations issues de la méthodologie présentée au chapitre 3.

Nous considérons les trois critères audio exprimés dans la partie 4.1.2.2 par un rapport de vraisemblance généralisé. Le critère morphologique n'étant actuellement pas formulé de la même manière, nous ne l'étudions pas dans l'approche multicritère. Chaque critère est considéré selon deux variantes, c'est-à-dire calculé sur les MFCCs ou les vecteurs de chroma. Les six variantes sont évaluées séparément dans le cadre de systèmes partiellement réglés en connaissant les frontières de référence (cadre oracle). Ces expériences permettent de choisir les variantes suivantes pour l'étude de l'approche multicritère : le critère de rupture d'homogénéité calculé sur les MFCCs, le critère

de rupture de répétition calculé sur les vecteurs de chroma et le critère de détection d'événements courts calculé sur les MFCCs. L'évaluation oracle du système combinant ces variantes à l'aide d'une somme pondérée permet d'obtenir de meilleures performances en comparaison de celles obtenues avec la meilleure variante seule, ce qui montre l'intérêt de notre approche.

Nous étudions ensuite l'apport lié à l'introduction d'une contrainte de régularité dans l'estimation des frontières structurelles. Nous considérons dans ce cadre un système combinant le meilleur critère suivant notre précédente étude et la contrainte de régularité dérivée de la valeur absolue à la puissance α. Nous calculons les performances du système pour un ensemble de réglages faisant varier l'importance de la contrainte devant le critère puis la convexité de cette contrainte. Nous observons une augmentation des performances pour plusieurs réglages en comparaison des performances oracles obtenues avec le critère seul, ce qui permet de mettre en avant l'utilisation d'une contrainte de régularité pour l'estimation des frontières structurelles. Le système considéré est enfin évalué dans un cadre réaliste à l'aide d'un processus de validation croisée.

Une évaluation d'un système combinant l'approche multicritère et la contrainte de régularité est enfin menée par validation croisée et permet d'observer des performances semblables à une version mono-critère de ce système. Ces résultats nous incitent à rechercher d'autres critères audio pertinents, des processus de combinaison de critères plus complexes ainsi que d'autres contraintes de régularité.

Dans un deuxième temps nous évaluons quatre variantes du système d'étiquetage introduit dans la partie 4.2 dans un but exploratoire. Les performances obtenues permettent de mettre en avant l'équivalence des deux critères de sélection d'automate considérés et montrent que l'utilisation du modèle système - contraste dans ce cadre est prometteuse. L'annotation sémiotique de la base RWC Pop étant actuellement en cours de finalisation, les présentes évaluations sont réalisées à l'aide d'une base de 100 morceaux annotés par l'IRCAM dans le cadre du projet Quaero.

Conclusion

Résumé

Les évolutions technologiques font qu'il est aujourd'hui facile d'accéder à de grandes bases de contenus musicaux, mais rendent impérieux le besoin de développer de nouvelles représentations et de nouveaux algorithmes pour les exploiter de manière optimale. Afin de pouvoir bénéficier d'une vision représentative de ces bases et de leurs éléments constitutifs, il est nécessaire de les caractériser efficacement à l'aide de descriptions pertinentes, comme leur structure. Dans cette thèse, nous nous sommes focalisés sur l'estimation de la structure macroscopique des morceaux de musique "conventionnels", c'est-à-dire diffusés par les médias de masse. Il s'agit de produire une description de leur organisation par une séquence de quelques dizaines de segments structurels étiquetés selon la ressemblance de leur contenu musical.

La diversité de la musique implique qu'il est particulièrement complexe de décrire cette organisation de manière unique, d'où l'absence de méthodologie communément admise constatée au début de cette thèse. Par le présent travail, nous avons contribué à la spécification et l'élaboration d'une méthodologie d'annotation d'une structure particulière : la structure sémiotique. Celle-ci s'appuie sur un ensemble de fondements et de concepts inspirés de la linguistique structuraliste, et ses principes sont formulés de sorte à couvrir un large éventail de genres et styles musicaux. L'analyse qui en résulte ne nécessite pas de connaissances musicologiques poussées de la part de l'annotateur, et la méthode est conçue pour que l'annotation soit aussi reproductible que possible. Ces considérations méthodologiques ont permis la création de plusieurs bases d'annotations utilisées dans le cadre de campagnes d'évaluation d'envergure nationale et internationale (Quaero et MIREX).

Du point de vue algorithmique, nous nous sommes tout d'abord concentrés sur l'estimation des frontières structurelles, en formulant le processus de segmentation comme l'optimisation d'un coût faisant apparaître clairement les contributions liées à la manière de caractériser les segments structurels (critères audio) et celles des *a priori* sur la structure visée (contraintes de régularité). À partir de cette formulation, nous avons exploré l'intérêt d'utiliser une contrainte sur les tailles de segments (pulsation structurelle) et l'apport de la combinaison de plusieurs critères audio par fusion. Nous avons évalué et diagnostiqué plusieurs systèmes basés sur ces approches, notamment dans le cadre de campagnes d'évaluation nationales et internationales. Ceci nous a permis de montrer que l'introduction d'une contrainte de régularité privilégiant une taille particulière de segments structurels est un facteur d'amélioration significatif, en particulier si une convention similaire est adoptée pour décrire la structure lors de la phase d'annotation. Par ailleurs, l'utilisation conjointe de trois critères audio, formulés dans un même cadre et combinés par une somme pondérée, permet d'obtenir des résultats du niveau de

l'état de l'art mais produit une progression modeste des performances en comparaison de l'utilisation du meilleur critère seul. L'approche par fusion reste cependant un axe de recherche prometteur.

Nous avons enfin élaboré un module d'estimation des étiquettes sémiotiques fondé sur la sélection d'automates probabilistes à états finis, évalué sous l'hypothèse de la connaissance préalable des frontières structurelles. L'évaluation de plusieurs configurations de ce module a permis d'une part d'introduire un critère auto-adaptatif expérimental de sélection de modèles donnant des performances équivalentes à celles d'un critère de l'état de l'art, et d'autre part de montrer l'intérêt de considérer un modèle d'organisation interne des segments structurels dans ce contexte.

Perspectives à court terme

Les résultats des travaux de cette thèse mettent également en lumière un certain nombre d'axes d'exploration et de marges de progression supplémentaires.

La contrainte de régularité est actuellement mise en oeuvre par l'intermédiaire d'un modèle simple favorisant une taille particulière de segment structurel, et dont les paramètres sont fixés *a priori*. Il nous semble intéressant d'étudier les manières d'adapter ces paramètres à chaque morceau à l'aide des critères audio déjà calculés ou en utilisant de nouveaux critères. notamment par l'intermédiaire des critères audio calculés. Nous estimons également qu'il peut être intéressant de rechercher d'autres modèles permettant de mieux prendre en compte les variations de taille des segments de même étiquette structurelle (extensions, troncatures), de telles variations étant fréquemment observées dans les morceaux de musique.

L'utilisation conjointe de plusieurs critères audio nous semble être une approche valable, qui peut progresser par l'amélioration des critères eux-mêmes. Nous utilisons actuellement un critère de rupture d'homogénéité ainsi qu'un critère de détection d'événements locaux basés sur l'analyse de l'évolution du timbre des morceaux, et un critère de rupture de répétition du contenu tonal. Le critère de rupture d'homogénéité donne des résultats bien meilleurs que les deux autres, et ce pour plusieurs raisons.

D'une part, le critère de détection d'événements courts présente en général un comportement erratique qui conduit à détecter un ensemble d'instants en surnombre par rapport à ceux qui correspondent à la fin des segments structurels. L'hypothèse de la présence d'événements courts marquant la fin des segments n'est donc pas très pertinente, et nous incite plutôt à considérer un modèle d'organisation interne des segments structurels (système - contraste). Il nous semble en outre prometteur de reformuler ce critère morphologique par un rapport de vraisemblance généralisé, comme c'est le cas pour les autres critères audio, afin de l'intégrer à notre approche. D'autre part, le critère de rupture de répétition n'est actuellement pas robuste aux variations de taille des segments de même étiquette, ni aux transpositions pouvant intervenir au cours du morceau. Ceci peut être pallié par l'utilisation d'une distance adéquate entre les séquences de descripteurs qui mesurerait leur alignement optimal parmi toutes les transpositions possibles. Enfin, un axe de progression intéressant consiste en la recherche de stratégies plus complexes de combinaison des critères, dans l'optique de les complémenter ou les corriger automatiquement.

Nos travaux ont privilégié les descripteurs musicaux caractérisant le timbre et le contenu tonal des morceaux, propriétés qui sont habituellement utilisées pour l'estimation de structure. Il est important de faire appel à d'autres dimensions musicales

dans l'optique de pouvoir exploiter celles qui sont les plus structurantes (contenu redondant, régularité de l'apparition de certains événements...) à l'échelle du morceau de musique. Ainsi, de nouveaux descripteurs méritent d'être ajoutés aux "classiques" MFCCs et vecteurs de chroma en vue de rendre compte de toutes les dimensions musicales. Mentionnons que l'utilisation d'outils de séparation de sources sonores peut également contribuer à mieux décrire individuellement ces différentes dimensions : par exemple, un extracteur de la partie vocale des chansons combinés avec un algorithme de transcription peut être utile pour modéliser l'information vocale.

Bien que ce point n'ait pu être exploré complètement au sein de la thèse, il est prometteur d'intégrer le modèle d'organisation interne des segments de la structure sémiotique dans le processus d'estimation des étiquettes structurelles : l'identification de la composante commune des segments associés à une même étiquette peut permettre une meilleure classification des segments structurels, robuste aux imprécisions sur l'estimation des frontières ainsi qu'à la présence d'affixes ou de troncatures.

Enfin, l'utilisation de grammaires relatives à l'agencement des segments structurels du morceau peut contribuer à améliorer l'estimation de leurs étiquettes sémiotiques. Ceci permettrait d'intégrer dans les approches algorithmiques l'analyse syntagmatique, laquelle constitue l'un des principes méthodologiques proposés pour l'annotation de la structure sémiotique.

Perspectives à long terme

Outre leur intérêt propre en description et estimation de la structure musicale, nos approches constituent également une contribution vers l'amélioration d'autres systèmes d'extraction de contenus musicaux.

D'un point de vue plus général, l'estimation d'une structure mono-dimensionnelle (dans le sens où les étiquettes structurelles n'ont qu'une dimension) est une première étape vers une modélisation multi-échelles et multidimensionnelle de l'organisation des morceaux de musique.

Nous nous sommes jusqu'à présent focalisés sur l'analyse de la structure sémiotique en considérant une échelle temporelle privilégiée. Cependant le contenu d'un morceau peut clairement être décrit simultanément à des échelles plus fines ou plus grossières : un nouveau pas consistera à intégrer cet aspect hiérarchique aux modèles ultérieurs de la structure sémiotique. Le fait de modéliser avec une résolution multiple l'organisation des morceaux de musique peut d'ailleurs renforcer la qualité de l'estimation de la structure sémiotique elle-même. Cette préoccupation rejoint plusieurs travaux récents qui considèrent une description hiérarchique de la structure, notamment ceux de Martin dans lesquels une recherche des répétitions du contenu tonal du morceau est effectuée à des échelles enchâssées les unes dans les autres pour en déduire une structure hiérarchique de répétitions [MHRF11].

Une modélisation multi-échelles et multidimensionnelle de l'organisation du contenu musical permet d'envisager la localisation temporelle et fréquentielle des composantes communes aux segments structurels de même étiquette (systèmes porteurs) et de les séparer de ce qui les distingue : il peut s'agir de variations locales (contraste), ou de variations observées à l'échelle du bloc. Ceci présente un intérêt notable pour une nouvelle représentation compacte des morceaux de musique, en tirant partie de la redondance entre les systèmes porteurs des segments correspondant à la même classe sémiotique. Cette modélisation peut contribuer à des applications de "stockage intelligent", de

représentation intuitive de contenus musicaux voire s'intégrer comme une des étapes de pré-traitement dans un système de transcription automatique de partitions.

Enfin, notons que le problème d'estimation de structure n'est pas limité au domaine de la musique. On peut imaginer à l'avenir adapter certaines de nos approches à la description structurelle d'autres contenus multimédia comme les flux radiophoniques, les images fixes, les vidéos, ou transposer certains concepts développés dans cette thèse à des domaines tels que la bio-informatique.

Annexes

Annexe A

Informations relatives aux bases de morceaux mentionnées dans la thèse

Liste des titres de la base Quaero

Quaero 2009 - Ensemble de développement (20 titres)			
Index	Titre	Artiste	Album
0009	09 Brain Damage	Pink Floyd	Dark Side of the Moon
0017	Lazing on a sunday afternoon	Queen	A Night at the Opera
0056	Mad Blunted Jazz	DJ Cam	Mad Blunted Jazz
0067	Return of the G	OutKast	Aquemini
0079	You Shook me all night long	ACDC	Back in Black
0090	Old Love	Eric Clapton	Unplugged
0099	O Pato	Stan Getz Juan Gilberto	Getz-Gilberto 2
0103	Caribbean blue	Enya	Shepherd Moons
0111	Off the wall	Mickael Jackson	Off the wall
0159	03 Planet	Bass America Collection 3	
0179	Fuk	Plastikman	Musik
0184	01 Natalies Party	Shack	HMS Fable
0222	Take You There	Sean Kingston	Sean Kingston
0231	Shawty Get Loose	Lil Mama	Voice of the Young People
0251	Waterloo	Abba	
0258	Blue (Da Ba Dee)	Eiffel 65	
0264	I d Do Anything For You (But I Won t Do T	Meat Loaf	
0278	Lambada	Kaoma	
0292	Conquest of Paradise	Vangelis	
0347	Smells Like Teen Spirit	Nirvana	

Quaero 2009 - Ensemble de test (49 titres)			
Index	Titre	Artiste	Album
0002	02 breathe	Pink Floyd	Dark Side of the Moon
0004	04 Time	Pink Floyd	Dark Side of the Moon
0007	07 Us and them	Pink Floyd	Dark Side of the Moon
0011	Cleanin Out my Closet	Eminem	The Eminem Show
0016	I m in love with my car	Queen	A Night at the Opera
0024	Little umbrellas	Franck Zappa	Hot Rats
0030	04 The Rage of Angels	Jedi Mind Tricks	Visions of Ghandi
0059	03 siamese twins	The Cure	Pornography
0060	04 the hanging garden	The Cure	Pornography
0061	05 the figurehead	The Cure	Pornography
0062	06 a strange day	The Cure	Pornography
0063	07 cold	The Cure	Pornography
0066	Cmon and love me	Kiss	Alive
0073	Brown Sugar	D Angelo	Brown Sugar
0078	Hells Bells	ACDC	Back in Black
0080	Cryin	Aerosmith	Get A Grip
0086	Layla	Eric Clapton	Unplugged
0089	Nobody Knows you when you're down	Eric Clapton	Unplugged
0093	Tears in heaven	Eric Clapton	Unplugged
0094	Walking blues	Eric Clapton	Unplugged
0104	Wedding Cocek	Goran Bregovic	Underground
0107	til death	Obituary	Slowly we rot
0112	Karma Police	Radiohead	OK Computer
0138	12 Polythene Pam	The Beatles	Abbey Road
0165	Jonz in my bonz	DAngelo	Brown Sugar
0168	Shit Damn Motherfucker	DAngelo	Brown Sugar
0183	Konception	Platiskman	Musik
0208	Take Me To The Fires	Waco Brothers	Cowboy In Flames
0211	we fly high	Jim Jones	Hustlers POME
0217	What Goes Around -	Justin Timberlake	
0224	Sorry	Buckcherry	15
0226	Love In This Club	Usher	Here I Stand
0229	01 Faith	Georges Michael	Faith
0230	Lollipop	Lil Wayne	Tha Carter III
0232	Touch my body	Mariah Carey	E MC2
0233	4 minute	Madonna	Hard Candy
0253	X	Xzibit	
0254	Believe	Cher	
0261	Dont worry Be happy	Bobby McFerrin	
0262	It's like	Run DMC	
0270	Without Me (Radio Edit)	Eminem	

Quaero 2009 - Ensemble de test (49 titres)			
Index	Titre	Artiste	Album
0275	Another Brick In The Wall	Pink Floyd	
0276	Kung Fu fighting	Carl Douglas	
0280	Born to be alive	Patrick Hernandez	
0281	I ll Be Missing You	Puff Daddy Feat Faith Evans	
0282	Fox on the run	Sweet	
0293	Winds of Change	Scorpions	
0301	Baby One More Time	Britney Spears	
0302	La Isla Bonita	Madonna	

Quaero 2010 (45 titres)			
Index	Titre	Artiste	Album
0005	05 Great Gig in the sky	Pink Floyd	Dark Side of the Moon
0010	10 Eclipse	Pink Floyd	Dark Side of the Moon
0012	39	Queen	A Night at the Opera
0014	Death on two legs	Queen	A Night at the Opera
0015	Good company	Queen	A Night at the Opera
0018	Love of my life	Queen	A Night at the Opera
0021	Sweet lady	Queen	A Night at the Opera
0028	02 Tibetan Black Magician	Jedi Mind Tricks	Visions of Ghandi
0034	08 A Storm of Words	Jedi Mind Tricks	Visions of Ghandi
0038	12 Rise of the Machines	Jedi Mind Tricks	Visions of Ghandi
0042	16 The Heart of Darkness	Jedi Mind Tricks	Visions of Ghandi
0049	Timbarma	Ali Farka Toure	Ali Farka Toure
0054	Natural Blues	Moby	Push
0055	Zombie	Cranberries	No Need to Argue
0057	01 100 years	The Cure	Pornography
0070	Safe From Harm	Massive Attack	Blue Lines
0075	Let love find you	Pucho and his latin soul brothers	The Best Of
0076	God Box	The Fall	458489 A Sides
0084	Before you accuse me	Eric Clapton	Unplugged
0091	Running on faith	Eric Clapton	Unplugged
0095	What's In it for me	Faith Hill	Breathe
0098	Tell me why	Neil Young	After the gold rush
0100	Beggin	Madcon	So Dark The Con Of Man

Quaero 2010 (45 titres)			
0105	You Dont know me	Ray Charles	Modern Sounds In Country And Western Music
0115	03 Baby s In Black	The Beatles	Beatles For Sale
0119	07 Kansas City Hey-Hey-Hey-Hey	The Beatles	Beatles For Sale
0123	11 Every Little Thing	The Beatles	Beatles For Sale
0127	01 Come Together	The Beatles	Abbey Road
0133	07 Here Comes The Sun	The Beatles	Abbey Road
0140	14 Golden Slumbers	The Beatles	Abbey Road
0145	02 I Should Have Known Better	The Beatles	A Hard Days Night
0150	07 Cant Buy Me Love	The Beatles	A Hard Days Night
0154	11 When I Get Home	The Beatles	A Hard Days Night
0158	02 Musiz-Electro V16	Bass America Collection 3	
0163	You know i'm no good	Amy Winehouse	Back to Black
0167	Me And Those Dreamin Eyes Of Min	DAngelo	Brown Sugar
0174	I Dream Of Jeannie Danny Boy	Joan Baez	Diamonds and Rust
0185	02 Comedy	Shack	HMS Fable
0186	01 Cocaine In My Brain	Dillinger	Cocaine
0214	its not over	Daughtry	Daughtry
0215	The Sweet Escape	Gwen Stefani	The Sweet Escape
0216	Runaway love	Ludacris	Release Therapy
0218	Low	Flo Rida Feat. TPain	
0219	Paralyzer	Finger Eleven	Them Vs You Vs Me
0223	New Soul	Yael Naim	Yael Naim and David Donatien

Quaero 2011 (45 titres)			
Index	Titre	Artiste	Album
0006	06 Money	Pink Floyd	Dark Side of the Moon
0008	08 Any color you like	Pink Floyd	Dark Side of the Moon
0020	Seaside rendez vous	Queen	A Night at the Opera
0026	Son of Mr Green genes	Franck Zappa	Hot Rats
0027	01 Intro	Jedi Mind Tricks	Visions of Ghandi
0029	03 Blood in Blood out	Jedi Mind Tricks	Visions of Ghandi

Quaero 2011 (45 titres)			
Index	Titre	Artiste	Album
0032	06 Animal Rap	Jedi Mind Tricks	Visions of Ghandi
0033	07 Nada Cambia	Jedi Mind Tricks	Visions of Ghandi
0036	10 The Wolf	Jedi Mind Tricks	Visions of Ghandi
0037	11 Walk with me	Jedi Mind Tricks	Visions of Ghandi
0039	13 Pity of War	Jedi Mind Tricks	Visions of Ghandi
0040	14 Kublain Khan	Jedi Mind Tricks	Visions of Ghandi
0041	15 Whats really good	Jedi Mind Tricks	Visions of Ghandi
0044	Bakoye	Ali Farka Toure	Ali Farka Toure
0045	Bakoytereye	Ali Farka Toure	Ali Farka Toure
0047	Nawiye	Ali Farka Toure	Ali Farka Toure
0051	Chan chan	Buenavista social club	Buenavista social club
0052	De camino a la vereda	Buenavista social club	Buenavista social club
0053	El Cuarto de Tula	Buenavista social club	Buenavista social club
0058	02 a short term effect	The Cure	Pornography
0064	08 pornography	The Cure	Pornography
0069	Travelling Man	Dolly Parton	Coat Of Many Colors
0072	Thug Love Remix	50 Cent Eminem Destiny Child	
0077	Back in Black	ACDC	Back in Black
0082	Its oh so quiet	Bjork	Post
0083	Alberta	Eric Clapton	Unplugged
0085	Hey Hey	Eric Clapton	Unplugged
0087	Lonely stranger	Eric Clapton	Unplugged
0088	Malted Milk	Eric Clapton	Unplugged
0097	Wanna be starting something	Mickael Jackson	Thriller
0101	Pull Together	Shack	HMS Fable
0102	Do You Want My Love	CoCo Lee	Just No other Way
0106	Breaking the law	Judas Priest	
0113	01 No Reply	The Beatles	Beatles For Sale
0116	04 Rock And Roll Music	The Beatles	Beatles For Sale
0117	05 I ll Follow The Sun	The Beatles	Beatles For Sale
0118	06 Mr	The Beatles	Beatles For Sale
0120	08 Eight Days A Week	The Beatles	Beatles For Sale
0121	09 Words Of Love	The Beatles	Beatles For Sale
0122	10 Honey Dont	The Beatles	Beatles For Sale
0125	13 What You re Doing	The Beatles	Beatles For Sale
0126	14 Everybody s Trying To Be My Baby	The Beatles	Beatles For Sale
0128	02 Something	The Beatles	Abbey Road
0130	04 Oh Darling	The Beatles	Abbey Road
0131	05 Octopus s Garden	The Beatles	Abbey Road

Annexe B

Annexes du chapitre 4

B.1 Détails du calcul de la forme analytique du critère de rupture d'homogénéité

Explicitons la formule du RVG dans la partie 4.1.2.2 dans le cas des descripteurs acoustiques, afin de la rendre analytique.

Sous hypothèse d'indépendance de la suite des variables aléatoires ayant produit y^0, on obtient

$$P(y^1|G_1) = \prod_{t=1}^{N} p(y_t^1|G_1). \tag{B.1}$$

Une observation x à valeurs dans $\mathbb{R}^d$, avec $d \in \mathbb{N}$, issue d'une distribution gaussienne de moyenne μ et de matrice de covariance Γ possède la vraisemblance suivante :

$$p(x|\mu,\Gamma) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det\Gamma}} \exp\left(-\frac{1}{2}(x-\mu)^T\Gamma^{-1}(x-\mu)\right). \tag{B.2}$$

Ainsi, en supposant que les éléments de y^1 sont issus d'une distribution gaussienne G_1 de moyenne μ_1 et de matrice de covariance Γ_1, on obtient

$$p(y^1|G_1) = \left(\frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det\Gamma}}\right)^N \exp\left(-\frac{1}{2}\sum_{t=1}^{N}(y_t-\mu_1)^T\Gamma_1^{-1}(y_t-\mu_1)\right). \tag{B.3}$$

Or

$$\frac{1}{N}\sum_{t=1}^{N}(y_t-\mu_1)^T\Gamma_1^{-1}(y_t-\mu_1)) = tr(\Gamma_1\Gamma_1^{-1}) = d. \tag{B.4}$$

D'où

$$p(y^1|G_1) = \left(\frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det\Gamma}}\right)^N \exp\left(-\frac{Nd}{2}\right). \tag{B.5}$$

On en déduit que

$$\log\left(p(y^1|G_1)\right) = -N\log\left((2\pi)^{\frac{d}{2}}\sqrt{\det\Gamma}\right) - \frac{Nd}{2} = -\frac{N}{2}\left(d(\log(2\pi)) + 1\right) + \log\left(\det(\Gamma_1)\right). \tag{B.6}$$

En procédant de même avec y^2 et y^0 en notant $G_2(\mu_2, \Gamma_2)$ et $G_0(\mu_0, \Gamma_0)$ leurs distributions gaussiennes respectives, on obtient :

$$\log(\text{RVG}) = -\frac{N}{2}(d(\log(2\pi))+1)+\log(\det(\Gamma_1))-\frac{N}{2}(d(\log(2\pi))+1)+\log(\det(\Gamma_2)) + N(d(\log(2\pi))+1)+\log(\det(\Gamma_0)). \quad \text{(B.7)}$$

Soit

$$\log(\text{RVG}) = N\left[\log(\det(\Gamma_0)) - \frac{\log(\det(\Gamma_1))+\log(\det(\Gamma_2))}{2}\right]. \quad \text{(B.8)}$$

Cette quantité est maximale lorsque y^1 et y^2 suivent des distributions gaussiennes distinctes.

B.2 Détails du calcul de la forme analytique du critère de rupture de répétition

Dans le cadre du calcul du RVG dans la partie 4.1.2.2 pour les descripteurs numériques, on suppose que Θ_0 correspond à l'espace des paramètres des lois de probabilité gaussienne à variance fixe, et que Θ_1 correspond à l'espace des paramètres des paires de lois de probabilité gaussienne à variance fixe, produisant toute sous-séquence de y de taille N. Ceci permet d'obtenir une expression analytique du log(RVG).

On a :

- $P(y^1|g^1) = \prod_{t=1}^{N} p(y_t^1|g_t^1)$ sous hypothèse d'indépendance des variables aléatoire du processus Y.
- $\log\left(p(y_t^1|g_t^1)\right) = -\rho_t\|g_t^1 - y_t^1\|^2 + \nu_t$, par hypothèse de modélisation gaussienne ($\|.\|$ représente la distance euclidienne)

D'où

$$\log\left(P(y^1|g^1)\right) = \sum_{t=1}^{N}(-\rho_t\|g_t^1 - y_t^1\|^2 + \nu_t). \quad \text{(B.9)}$$

En supposant l'égalité des variances, $\rho_t = \lambda$ on obtient

$$\log\left(P(y^1|g^1)\right) = -\rho_t\sum_{t=1}^{N}(\|g_t^1 - y_t^1\|^2) + \sum_{t=1}^{N}\nu_t. \quad \text{(B.10)}$$

En procédant de la même manière pour $P(y^0|g^0)$ et $P(y^2|g^2)$, on obtient le critère de répétition suivant :

$$\log(\text{RVG}) \propto -\sum_{t=1}^{N}\|g_t^1 - y_t^1\|^2 - \sum_{t=1}^{N}\|g_t^2 - y_t^2\|^2 + \sum_{t=1}^{2N}\|g_t^0 - y_t^0\|^2 + \textit{constante}. \quad \text{(B.11)}$$

Bibliographie

[AdM01] Claude Abromont and Eugène de Montalembert. *Guide de la théorie de la musique*. Éditions Fayard / Henry Lemoine, 2001.

[AdM10] Claude Abromont and Eugène de Montalembert. *Guide des formes de la musique occidentale*. Éditions Fayard / Henry Lemoine, 2010.

[AHdS96] Phipps Arabie, Lawrence J. Hubert, and Geert de Soete, editors. *Clustering and classification*. World Scientific Publishing, 1996.

[Aka74] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6) :716–723, December 1974.

[ANO06] Kamil Adiloglu, Thomas Noll, and Klaus Obermayer. A paradigmatic approach to extract the melodic structure of a musical piece. *Journal of New Music Research*, 35(3) :221–236, 2006.

[ANS+05] Samer Abdallah, Katy Noland, Mark Sandler, Michael Casey, and Christophe Rhodes. Theory and evaluation of a bayesian music structure extractor. In *Proceedings of the 6th International Conference on Music Information Retrieval*, pages 420–425, September 2005.

[AS01] Jean-Julien Aucouturier and Mark Sandler. Segmentation of musical signals using hidden markov models. In *Proceedings of the Audio Engineering Society (AES) 110th Convention*, May 2001.

[AS02] Jean-Julien Aucouturier and Mark Sandler. Finding repeating patterns in acoustic musical signals : Applications for audio thumbnailing. In *Proceedings of the AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, pages 412–421, Espoo, Finand, June 2002.

[BCL10] Luke Barrington, Antoni B. Chan, and Gert Lanckriet. Modeling music as a dynamic texture. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3) :602–612, March 2010.

[BD98] Ian Bent and William Drabkin. *Analysis*. Macmillian reference limited, 1988 (reprinted 1998).

[BDSV11] Frédéric Bimbot, Emmanuel Deruty, Gabriel Sargent, and Emmanuel Vincent. Methodology and resources for the structural segmentation of music pieces into autonomous and comparable blocks. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 287–292, October 2011.

[BDSV12a] Frédéric Bimbot, Emmanuel Deruty, Gabriel Sargent, and Emmanuel Vincent. Methodology and conventions for the latent semiotic annotation of music structure. Report PI-1993, IRISA, February 2012.

[BDSV12b] Frédéric Bimbot, Emmanuel Deruty, Gabriel Sargent, and Emmanuel Vincent. Semiotic Structure Labeling of Music Pieces : Concepts, Methods and Annotation Conventions . In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 235–240, October 2012.

[BDSV12c] Frédéric Bimbot, Emmanuel Deruty, Gabriel Sargent, and Emmanuel Vincent. System & contrast : a polymorphous model of the inner organization of structural segments within music pieces. Report PI 1999, IRISA, December 2012.

[Bel54] Richard Bellman. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, (60) :503–516, 1954.

[BLSV10a] Frédéric Bimbot, Olivier Le Blouch, Gabriel Sargent, and Emmanuel Vincent. Décomposition en blocs autonomes comparables - Une proposition de description et d'annotation de structure pour le traitement automatique des morceaux de musique. In *Actes des Xèmes Journées d'Informatique Musicale (JIM)*, May 2010.

[BLSV10b] Frédéric Bimbot, Olivier Le Blouch, Gabriel Sargent, and Emmanuel Vincent. Decomposition into autonomous and comparable blocks : A structural description of music pieces. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 189–194, August 2010.

[BMK06] Michael J. Bruderer, Martin McKinney, and Armin Kohlrausch. Structural boundary perception in popular music. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*, pages 198–201, October 2006.

[BMM95] Frédéric Bimbot, Ivan Magrin-Chagnolleau, and Luc Mathan. Second-order statistical measures for text-independant speaker identification. *Speech Communication*, 17(1-2) :177–192, August 1995.

[BMTP99] Emmanuel Bigand, François Madurell, Barbara Tillmann, and Marion Pineau. Effect of global structure and temporal organization on chord processing. *Journal of Experimental Psychology : Human Perception and Performance*, 25(1) :184–197, 1999.

[BW01] Mark A. Bartsch and Gregory H. Wakefield. To catch a chorus : Chroma-based representations for audio thumbnailing. In *Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 15–18, October 2001.

[BW05] Mark A. Bartsch and Gregory H. Wakefield. Audio thumbnailing of popular music using chroma-based representations. In *IEEE Transactions on multimedia*, volume 7, pages 96–104, February 2005.

[Cap98] William E. Caplin. *Classical Form*. Oxford University Press, 1998.

[CL11a] Ruofeng Chen and Ming Li. Music structural segmentation by combining harmonic and timbral information. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 477–482, Oct 2011.

[CL11b] Ruofeng Chen and Ming Li. Music structural segmentation by combining harmonic and timbral information (MIREX 2011). In *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, october 2011.

[CVG+08] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and Malcom Slaney. Content-based music information retrieval : Current directions and future challenges. *Proceedings of the IEEE*, 96(4) :668–696, April 2008.

[Dan05] Roger Dannenberg. Toward automated holistic beat tracking, music analysis and understanding. In *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)*, pages 366–373, September 2005.

[Dav07] Matthew E.P. Davies. *Towards Automatic Rhythmic Accompaniment*. PhD thesis, Department of Electronic Engineering, Queen Mary, University of London, 2007.

[DBC09] J. Stephen Downie, Donald Byrd, and Tim Crawford. Ten years of ISMIR : Reflections on challenges and opportunities. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 13–18, October 2009.

[DG08] Roger Dannenberg and Masataka Goto. Music structure analysis from acoustic signals. In David Havelock, Sonoko Kuwano, and Michael Vörlander, editors, *Handbook of Signal Processing in Acoustics*, volume 1, pages 305–331. Springer, 2008.

[DH02] Roger Dannenberg and Ning Hu. Music structural segmentation by combining harmonic and timbral information. In *Proceedings of the 3rd International Society for Music Information Retrieval Conference (ISMIR)*, October 2002.

[dS16] Ferdinand de Saussure. *Cours de Linguistique Générale*. 1916.

[EBD+11] Andreas Ehmann, Mert Bay, J. Stephen Downie, Ichiro Fujinaga, and David De Roure. Music structure segmentation algorithm evaluation : expanding on MIREX 2010 analyses and datasets . In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 561–566, Miami, United States, October 2011.

[Ehr05] Matthias Ehrgott. *Multicriteria optimization, second edition*. Springer, 2005.

[EK10] Antti Eronen and Anssi Klapuri. Music tempo estimation with k-NN regression. *IEEE transactions on Audio, Speech, and Language Processing*, 18(1) :50–57, 2010.

[EP07] Daniel P.W. Ellis and Graham E. Poliner. Identifying "cover songs" with chroma features and dynamic programming beat tracking. In *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages IV–1429–1432, April 2007.

[Fan60] Gunnar Fant. *Acoustic theory of speech production*. Mouton, The Hague, 1960.

[FC03] Jonathan T. Foote and Matthew L. Cooper. Media segmentation using self-similarity decomposition. In *Proceedings of the SPIE Storage and Retrieval for Multimedia Databases*, pages 167–175, January 2003.

[Foo00] Jonathan T. Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the 2000 IEEE International Conference on Multimedia and Expo (ICME)*, pages 452–455, August 2000.

[GHNO02] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database : Popular, classical, and jazz music databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*, pages 287–288, October 2002.

[Góm06] Emilia Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, Universitat Pompeu Fabra, Barcelona, 2006.

[Got03] Masataka Goto. A chorus-section detecting method for musical audio signals. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 437–440, April 2003.

[Gra84] Robert Gray. Vector quantization. *IEEE Acoustics, Speech, and Signal Processing Magazine*, pages 4–29, April 1984.

[GSMA12] Peter Grosche, Joan Serrà, Meinard Müller, and Josep Lluis Arcos. Structure-based audio fingerprinting for music retrieval. In *Proceedings of the 13th International Society on Music Information Retrieval (ISMIR)*, pages 55–60, October 2012.

[Hje43] Louis Hjelmslev. *Prolégomènes à une théorie du langage*. 1943.

[Jeh05] Tristan Jehan. Hierarchical multi-class self similarities. In *Proceedings of the 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 311–314, New Paltz, New York, October 2005.

[Jen07] Kristoffer Jensen. Multiple scale music segmentation using rhythm, timbre and harmony. *EURASIP Journal on Advances in Signal Processing*, 2007 :1–11, 2007.

[KD06] Anssi Klapuri and Manuel Davy. *Signal Processing Methods for Music Transcription*. Springer, 2006.

[KS10] Florian Kaiser and Thomas Sikora. Music structure discovery in popular music using non-negative matrix factorization. *Proceedings of the 11th International Society on Music Information Retrieval (ISMIR)*, pages 429–434, October 2010.

[KSG12] Florian Kaiser, Thomas Sikora, and Peeters Geoffroy. MIREX 2012 - Music structural segmentation task : IRCAMstructure submission. In *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, 2012.

[LA11] Hélène Lachambre and Régine André-Obrecht. Baseline evaluation report on T6.5 : Music structuring and summarization. Technical report, IRIT, October 2011. Internal report of the Quaero Project.

[LC00] Beth Logan and Sephen Chu. Music summarization using key phrases. In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 749–752, June 2000.

[LNS07] Mark Levy, Katy Noland, and Mark Sandler. A comparison of timbral and harmonic music segmentation algorithms. In *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages IV–1433–1436, 2007.

[Log00] Beth Logan. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the 2nd Internantional Symposium on Music Information Retrieval (ISMIR)*, 2000.

[LPAK09] Ju-Hong Lee, Sun Park, Chan-Min Ahn, and Daeho Kim. Automatic generic document summarization based on non-negative matrix factorization. *Information Processing and Management*, (45) :20–34, 2009.

[LRB+12] Antoine Liutkus, Zafar Rafii, Roland Badeau, Bryan Pardo, and Gaël Richard. Adaptive filtering for musicvoice separation exploiting the repeating musical structure. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012.

[LS06] Mark Levy and Mark Sandler. New methods in structural segmentation of musical audio. In *Proceedings of European Signal Processing Conference (EUSIPCO)*, pages 318–326, September 2006.

[LS08] Mark Levy and Mark Sandler. Structural segmentation of musical audio by constrained clustering. *IEEE transactions on Audio, Speech and Language Processing*, 16(2) :318–326, February 2008.

[LWZ04] Lie Lu, Muyuan Wang, and Hong-Jiang Zhang. Repeating pattern discovery and structure analysis from acoustic music data. In *Proceedings of the 2004 Multimedia Information Retrieval Workshop*, pages 275–282, October 2004.

[MA04] R. Timothy Marler and Jasbir S. Arora. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization*, 26 :369–395, 2004.

[Mad06] Namunu C. Maddage. Automatic structure detection for popular music. *IEEE MultiMedia*, 13 :65–77, 2006.

[ME11] Meinard Müller and Sebastian Ewert. Chroma Toolbox : MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, pages 215–220, October 2011.

[MEK09] Meinard Müller, Sebastian Ewert, and Sebastian Kreuzer. Making chroma features more robust to timbre changes. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1877–1880, April 2009.

[MGJ11] Meinard Müller, Peter Grosche, and Nanzhu Jiang. A segment-based fitness measure for capturing repetitive structures of music recordings. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 615–620, 2011.

[MHRF11] Benjamin Martin, Pierre Hanna, Matthias Robine, and Pascal Ferraro. Structural analysis of harmonic features using string matching techniques (extended abstract). In *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, 2011.

[Mid90] Richard Middleton. *Studying popular music*. Open University Press, 1990.

[Mil74] Rupert G. Miller. The jackknife – a review. *Biometrika*, 61(1) :1–15, April 1974.

[MM04] Martin F. McKinney and Dirk Moelants. Tempo perception and musical content : what makes a piece fast, slow or temporally ambiguous ? In *Proceedings of the 8th International Conference on Music Perception and Cognition (ICMPC)*, pages 558–562, 2004.

[MND09] Matthias Mauch, Katy Noland, and Simon Dixon. Using musical structure to enhance automatic chord transcription. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 231–236, 2009.

[MRR80] Cory S. Myers, Laurence R. Rabiner, and Aaron E. Rosenberg. An investigation of the use of dynamic time warping for word spotting and connected speech recognition. pages 173–177, April 1980.

[Nat87] Jean-Jaques Nattiez. *Musicologie générale et sémiologie*. Christian Bourgois Editeur, 1987.

[NWL+07] Patrick Naïm, Pierre-Henri Wuillemin, Philippe Leray, Olivier Pourret, and Anna Becker. *Les réseaux bayésiens, 3e édition*. Eyrolles, 2007.

[Ori06] Nicola Orio. Music retrieval : A tutorial and review. In *Foundations and Trends in Information Retrieval*, pages 1–90, 2006.

[PD09] Geoffroy Peeters and Emmanuel Deruty. Is music structure annotation multi-dimensional ? A proposal for robust local music annotation. In *Proceedings of the 3rd International Workshop on Learning Semantics of Audio Signals (LSAS)*, pages 75–90, December 2009.

[PE05] Graham E. Poliner and Daniel P.W. Ellis. A classification approach to melody transcription. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, pages 161–166, September 2005.

[Pee07] Geoffroy Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum likelihood approach. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, pages 35–40, September 2007.

[Pee10] Geoffroy Peeters. MIREX 2010 music structure segmentation task : IRCAMsummary submission. In *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, October 2010.

[Pee11] Geoffroy Peeters. Music structure discovery : measuring the "state-ness" of times. In *Late-Breaking News from the 12th International Symposium for Music Information Retrieval (ISMIR)*, 2011.

[Pei07] Ewald Peiszer. Automatic audio segmentation : segment boundary and structure detection in popular music. Master's thesis, Vienna University of Technology, Austria, August 2007.

[PK06] Jouni Paulus and Anssi Klapuri. Music structure analysis by finding repeated parts. In *Proceedings of the 1st Audio and Music Computing (AMC) for Multimedia Workshop*, pages 59–68, October 2006.

[PK08a] J. Paulus and Anssi Klapuri. Music structure analysis using a probabilistic fitness measure and an integrated musicological model. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, pages 369–374, September 2008.

[PK08b] Jouni Paulus and Anssi Klapuri. Labelling the structural parts of a music piece with markov models. In *Proceedings of the 2008 Computers in Music Modeling and Retrieval Conference (CMMR)*, pages 137–147, May 2008.

[PKA11] Yannis Panagakis, Constantine Kotropoulos, and Gonzalo Arce. l1-graph based music structure analysis. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 495–500, October 2011.

[PLR02] Geoffroy Peeters, Amaury La Burthe, and Xavier Rodet. Toward automatic music audio summary generation from signal analysis. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*, pages 94–100, October 2002.

[PMK10] Jouni Paulus, Meinard Müller, and Anssi Klapuri. Audio-based music structure analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 625–636, August 2010.

[PP09] Hélène Papadopoulos and Geoffroy Peeters. Local key estimation based on harmonic and metric structures. In *Proceedings of the 12th International Conference on Digital Audio Effects (DaFX)*, pages 1–8, September 2009.

[Rab89] Laurence R. Rabiner. A tutorial on HMM and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286, February 1989.

[RC07] Christophe Rhodes and Michael. A. Casey. Algorithms for determining and labelling approximate hierarchical self-similarity. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, pages 41–46, September 2007.

[RCAS06] Christophe Rhodes, Michael A. Casey, Samer Abdallah, and Mark Sandler. A Markov-chain Monte-Carlo approach to musical audio segmentation. *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 797–800, May 2006.

[RJ93] Laurence R. Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice Hall Signal Processing Series, 1993.

[Ruw87] Nicolas Ruwet. Methods of analysis in musicology. *Music Analysis*, 6(1/2) :3–9+11–36, 1987.

[SBB00] Mouhamadou Seck, Raphaël Blouet, and Frédéric Bimbot. The IRISA/ELISA Speaker Detection and Tracking Systems for the NIST'99 Evaluation Campaign. *Digital Signal Processing*, 10(1-3) :154–171, January 2000.

[SBF+11] Jordan Bennett Louis Smith, J. Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen Downie. Design and creation of a large-scale database of structural annotations. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 555–560, October 2011.

[SBV10a] Gabriel Sargent, Frédéric Bimbot, and Emmanuel Vincent. A structural segmentation of songs using generalized likelihood ratio under regularity assumptions (extended abstract). In *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, october 2010.

[SBV10b] Gabriel Sargent, Frédéric Bimbot, and Emmanuel Vincent. Un système de détection de rupture de timbre pour la description de la structure des morceaux de musique. In *Actes des Journées d'informatique musicale (JIM)*, 2010.

[Sch52] Pierre Schaeffer. *A la recherche d'une musique concrète*. Seuil, 1952.

[Sch94] Véronique Schott-Bourget. *Approches de la linguistique*. Nathan Université, 1994.

[SDP09] Adam M. Stark, Matthew E. P. Davies, and Mark D. Plumbley. Real-time beat-synchronous analysis of musical audio. In *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx)*, September 2009.

[She64] Roger N. Shepard. Circularity in judgments of relative pitch. *Journal of the Acoustical Society of America*, 36(12) :2346–2353, 1964.

[Sir09] Jaques Siron. *Bases des mots aux sons*. Outre mesure, 2009.

[SJJ06] Yu Shiu, Hong Jeong, and C.-C. Jay Kuo. Similarity matrix processing for music structure analysis. In *Proceedings of the 14th ACM International Conference on Multimedia*, pages 69–76, October 2006.

[Slo91] John A. Sloboda. Music structure and emotional response : some empirical findings. *Psychology of Music*, 19 :110–120, 1991.

[SMPA12] Joan Serrà, Meinard Müller, Grosche Peter, and Josep Ll. Arcos. The importance of detecting boundaries in music structure annotation. In *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, oct 2012.

[Sny00] Bob Snyder. *Music and memory : an introduction*. MIT Press, 2000.

[SRS08] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka. Sparse and shift-invariant feature extraction from non-negative data. *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2069–2072, 2008.

[SZ10] Yu Shun-Zheng. Hidden semi-markov models. *Artificial Intelligence*, 174(2) :215–243, 2010.

[TC02] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 10(5) :293–302, July 2002.

[TLPG07] Douglas Turnbull, Gert Lanckriet, Elias Pampalk, and Masataka Goto. A supervised approach for detecting boundaries in music using difference features and boosting. *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, pages 51–54, 2007.

[TSB05] Hiroko Terasawa, Malcolm Slaney, and Jonathan Berger. The thirteen colors of timbre. In *Proceedings of the 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 323–326, October 2005.

[TWV05] Rainer Typke, Frans Wiering, and Remco C. Veltkamp. A survey on music information retrieval systems. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, pages 153–160, September 2005.

[UUN+10] Yushi Ueda, Yuki Uchiyama, Takuya Nishimoto, Nobutaka Ono, and Shigeki Sagayama. HMM-based approach for automatic chord detection using refined acoustic features. In *Proceedings of the 2010 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5506–5509, March 2010.

[vR79] Cornelis Joost van Rijsbergen. *Information Retrieval (2nd ed.)*. Butterworth, 1979.

[WB10] Ron Weiss and Juan Bello. Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization. In *Proceedings of the 11th International Society on Music Information Retrieval (ISMIR)*, pages 123–128, October 2010.

[Wei99] Yair Weiss. Segmentation using eigenvectors : a unifying view. In *Proceedings of the 1999 IEEE International Conference on Computer Vision (ICCV)*, pages 975–982, 1999.

Abstract

Recent progress in information and communication technologies makes it easier to access large collections of digitized music. New representations and algorithms must be developed in order to get a representative overview of these collections, and to browse their content efficiently. It is therefore necessary to characterize music pieces through relevant macroscopic descriptions. In this thesis, we focus on the estimation of the structure of music pieces : the goal is to produce for each piece a description of its organization by means of a sequence of a few dozen structural segments, each of them defined by its boundaries (starting time and ending time) and a label reflecting its audio content.

The notion of music structure corresponds to a wide range of meanings depending on the musical properties and the temporal scale under consideration. We introduce an annotation methodology based on the concept of "semiotic structure" which covers a large variety of musical styles. Structural segments are determined through the analysis of their similarities within the music piece, the coherence of their inner organization ("system-contrast" model) and their contextual relationship. A corpus of 383 pieces has been annotated according to this methodology and released to the scientific community.

In terms of algorithmic contributions, this thesis concentrates in the first place on the estimation of structural boundaries. We formulate the segmentation process as the optimization of a cost function which is composed of two terms. The first one corresponds to the characterization of structural segments by means of audio criteria. The second one relies on the regularity of the target structure with respect to a "structural pulsation period". In this context, we compare several regularity constraints and study the combination of audio criteria through fusion.

Secondly, we consider the estimation of structural labels as a probabilistic finite-state automaton selection process : in this scope, we propose an auto-adaptive criterion for model selection, applied to a description of the tonal content. We also propose a labeling method derived from the system-contrast model.

We evaluate several systems for structural segmentation of music based on these approaches in the context of national and international evaluation campaigns (Quaero, MIREX). Additional diagnostic is finally presented to complement this work.

Résumé

Les récentes évolutions des technologies de l'information et de la communication font qu'il est aujourd'hui facile de consulter des catalogues de morceaux de musique conséquents. De nouvelles représentations et de nouveaux algorithmes doivent de ce fait être développés afin de disposer d'une vision représentative de ces catalogues et de naviguer avec agilité dans leurs contenus. Ceci nécessite une caractérisation efficace des morceaux de musique par l'intermédiaire de descriptions macroscopiques pertinentes.

Dans cette thèse, nous nous focalisons sur l'estimation de la structure des morceaux de musique : il s'agit de produire pour chaque morceau une description de son organisation par une séquence de quelques dizaines de segments structurels, définis par leurs frontières (un instant de début et un instant de fin) et par une étiquette représentant leur contenu sonore.

La notion de structure musicale peut correspondre à de multiples acceptions selon les propriétés musicales choisies et l'échelle temporelle considérée. Nous introduisons le concept de structure "sémiotique" qui permet de définir une méthodologie d'annotation couvrant un vaste ensemble de styles musicaux. La détermination des segments structurels est fondée sur l'analyse des similarités entre segments au sein du morceau, sur la cohérence de leur organisation interne (modèle "système-contraste") et sur les relations contextuelles qu'ils entretiennent les uns avec les autres. Un corpus de 383 morceaux a été annoté selon cette méthodologie et mis à disposition de la communauté scientifique.

En termes de contributions algorithmiques, cette thèse se concentre en premier lieu sur l'estimation des frontières structurelles, en formulant le processus de segmentation comme l'optimisation d'un coût composé de deux termes : le premier correspond à la caractérisation des segments structurels par des critères audio et le second reflète la régularité de la structure obtenue en référence à une "pulsation structurelle". Dans le cadre de cette formulation, nous comparons plusieurs contraintes de régularité et nous étudions la combinaison de critères audio par fusion.

L'estimation des étiquettes structurelles est pour sa part abordée sous l'angle d'un processus de sélection d'automates à états finis : nous proposons un critère autoadaptatif de sélection de modèles probabilistes que nous appliquons à une description du contenu tonal. Nous présentons également une méthode d'étiquetage des segments dérivée du modèle système-contraste.

Nous évaluons différents systèmes d'estimation automatique de structure musicale basés sur ces approches dans le cadre de campagnes d'évaluation nationales et internationales (Quaero, MIREX), et nous complétons cette étude par quelques éléments de diagnostic additionnels.